UNIVERSIDADE D
COIMBRA

Maria Manuela de Sousa Freire

REDES SOCIAIS NO APOIO À DECISÃO

UM MODELO DE ANÁLISE PARA CONTEXTOS ESPECÍFICOS

Tese no âmbito do Doutoramento em Gestão - Ciência Aplicada à Decisão, orientada pelo Professor Doutor João Paulo Faria de Oliveira e Costa e pelo Professor Doutor Francisco José Peixeiro Antunes e apresentada à Faculdade de Economia da Universidade de Coimbra.

Julho de 2020

Maria Manuela de Sousa Freire

Redes sociais no apoio à decisão

Um modelo de análise para contextos específicos

Tese de Doutoramento em Gestão - Ciência Aplicada à Decisão, apresentada à Faculdade de Economia da Universidade de Coimbra para obtenção do grau de Doutor

Orientadores: Professor Doutor João Paulo Faria de Oliveira e Costa e Professor Doutor Francisco José Peixeiro Antunes

Coimbra, Julho de 2020

Ao André, Ivo e Zé Manel

Agradecimentos

Acabar este trabalho de investigação foi, por si só, o fechar de um ciclo e o início de outro. Durante este percurso cruzei-me com algumas pessoas que foram imprescindíveis para a sua concretização. Assim, quero aqui deixar o meu agradecimento e reconhecimento pelo apoio e disponibilidade que estes me deram.

Um especial obrigado aos meus orientadores Professor Doutor João Paulo Costa e ao Professor Doutor Francisco Antunes, pelo apoio, disponibilidade e motivação. Pelo conhecimento transmitido, conselhos e incentivos, bem como pela confiança que depositaram em mim. Obrigada por tudo o que aprendi convosco e pela paciência que ambos tiveram comigo para ultrapassar os obstáculos e seguir em frente.

Ao André, ao Ivo e ao Zé Manel por serem as pessoas incríveis que são, por me incentivarem e nunca me deixarem desistir, pelo seu apoio e força incondicionais.

Aos colegas de doutoramento, que partilharam este percurso, pelo apoio, estímulo e pelos momentos de descontração que proporcionaram.

Agradeço aos meus amigos e amigas por me incentivarem nesta missão, dando sempre uma palavra de estímulo nos momentos em que mais precisei.

Por último, o meu sincero agradecimento a todos(as) aqueles(as) que direta ou indiretamente contribuíram para que este trabalho de investigação chegasse a bom termo.

Coimbra, julho de 2020

We have a deeply rooted need to share our ideas and experiences, which gives us the ability to connect with other people, to be heard, and to feel a sense of worth and importance. We are curious about the world around us and how to organize and manipulate it, and we use communication to share our observations, ask questions, and engage with other people in meaningful dialogues about our quandaries.

(Russell e Klassen, 2019)

Resumo

As redes sociais *online* (RSO) possibilitam, sem inquirir ninguém, recolher informação das opiniões e experiências de um vasto conjunto de pessoas. Dado que a comunicação e a interação social são importantes para apoio à decisão, a análise das RSO e dos seus conteúdos desempenham um papel importante em todo esse processo. Os dados produzidos nas RSO são de uma riqueza tal, que permitem compreender o comportamento *online* e monitorizar tendências sociais para o apoio à decisão. Compreender as interações sociais e como podem apoiar a tomada de decisões representou a principal motivação deste trabalho. Nele foram focados os aspetos importantes da teoria dos grafos e das técnicas da *social network analysis* (SNA), bem como a sua utilização para apoio à decisão. Propôs-se um modelo para apoiar e otimizar as etapas do processo de análise de dados das RSO. A sua estrutura divide-se em três fases: (1) extração de dados; (2) processamento de dados; (3) interpretação e processamento semântico. A implementação do modelo foi feita com recurso a *software* de utilização comum, técnicas da SNA para visualização das redes e cálculo de métricas e técnicas de processamento de linguagem natural (PLN), tais como *data-mining* e *text-mining*, *data cleaning*, entre outras. O modelo integrou uma *graph database* que organizou os dados em formato grafo e para explorar e estudar as RSO combinaram-se duas perspetivas diferentes: (1) as interações sociais, que dependem das ligações entre utilizadores que difundem a informação; (2) os diálogos do discurso *web* baseados na análise semântica dos textos, que dependem do conteúdo da informação que é transmitida. Com a conjugação destas duas perspetivas analisaram-se os três níveis que caraterizam o discurso *web*: (1) as interações entre utilizadores (*user* vs *user*); (2) utilizado e mensagem(s) (*user* vs *post*); (3) mensagem(s) e conceito(s) (*post* vs *concept*). O objetivo principal foi ligar uma mensagem (*post*) a um utilizador e, por sua vez, um conceito a um *post*, para saber quem disse o quê. A análise dos dados semânticos teve como finalidade responder às questões e aos problemas de decisão de um contexto organizacional específico. Foram usados dados reais recolhidos do *Facebook*, em quatro estudos de caso, para mostrar a aplicabilidade do modelo, aos quais se aplicaram técnicas da SNA. Em conclusão, os resultados obtidos reforçam que, a extração de informação e conhecimento das interações entre utilizadores de RSO é importante para o apoio à decisão.

Palavras-chave: redes sociais online, *social network analysis*, apoio à decisão, discurso *web*, análise semântica, análise de dados.

Abstract

Online social networks (OSN) have made possible, without asking anyone, to collect information on the opinions and experiences of a wide range of people. Given that communication and social interaction are important for decision support, the analysis of OSN and its contents play an important role in this whole process. Data produced within OSN possess such richness that they make it possible to understand online behavior and monitor social trends for decision support. Understanding social interactions and how they can support decision making was the main motivation of this work. It focused on the important aspects of graph theory and social network analysis (SNA) techniques, as well as its use to support decision-making. A conceptual model was proposed to support and optimize the stages of OSN data analysis process. Its structure is divided into three phases: (1) data extraction; (2) data processing; (3) semantic interpretation and processing. The model was implemented using common software, SNA techniques for visualizing networks and calculating metrics and Natural Language Processing (NLP) techniques, such as data-mining and text-mining, data cleaning, among others. The model integrated a graph database which organized data in graph format and, to explore and study OSN, combining two different perspectives: (1) social interactions, which depend on the connections between users who disseminate the information; (2) the dialogues of the web discourse based on the semantic analysis of the texts, which depend on the content of the information that is transmitted. With the combination of these two perspectives, the three levels that characterize the web discourse were analyzed: (1) the interactions between users (user vs user); (2) user vs post; (3) post vs concept. The main objective was to associate a message (post) with a user and, in turn, a concept with a post, to determine who said what. The analysis of semantic data aimed to answer questions and decision problems of specific organizational contexts. Actual data collected from Facebook, in four case studies, were used to show the applicability of the model, to which SNA techniques were applied. In conclusion, the obtained results reinforce that extraction of information and useful knowledge from interactions between users of OSN are important to decision support.

Keywords: online social networks, social network analysis, decision support, web discourse, semantic analysis, data analysis.

Abreviaturas

API	*Application programming interface*
ASCII	*American standard code for information interchange*
CMC	Comunicação mediada por computador
CMCMD	*Convergent media computer mediated discourse*
XML	*eXtensible markup language*
CSV	*Comma-separated values*
JSON	*JavaScript object notation*
NLP	*Natural language processing*
NLTK	*Natural language toolkit*
OWL	*Web ontology language*
PLN	Processamento de linguagem natural
RDF	*Resouce description framework*
RGPD	Regulamento geral de proteção de dados
RSO	Rede social online
SAD	Sistemas de apoio à decisão
SI	Sistemas de informação
SNA	*Social network analysis*
TAP	Transportes aéreos portugueses
TI	Tecnologias de informação
TSV	*Tab-separated values*
VBA	*Visual basic for applications*
VW	Volkswagen
W3C	*World Wide Web Consortium*

Figuras

Tabelas

Sumário

1 Introdução

As tecnologias de informação (TI) aumentaram, quase de forma ilimitada, a capacidade de comunicarmos, informarmos e sermos informados, bem como de adquirirmos conhecimento (Fu et al., 2017; Savic et al., 2019). As TI dependem dos sistemas de informação (SI) e caracterizam-se por serem tecnologias digitais que, de acordo com Piedrahita et al. (2017), transformaram as redes interpessoais em estruturas compactas e densas que produzem informação constantemente. Essas redes interpessoais, no contexto das tecnologias digitais e da *web* social, designam-se de redes sociais *online* (RSO) e são cada vez mais utilizadas pelas empresas como fonte de informação.

As RSO são cada vez mais dinâmicas visto que são alimentadas pelo acesso democratizado às TI e pela elevada mobilidade geográfica (Fu et al., 2017). Para além dos conteúdos produzidos nessas redes se difundirem de forma muito expedita, estes podem ser utilizados, comentados e/ou criticados. Isto faz com que as RSO sejam plataformas importantes para disseminar informação, partilhar opiniões e criar discussões e debates sobre um qualquer tema. Se por um lado temos utilizadores mais esclarecidos, mais ativos e mais críticos, por outro, temos empresas que divulgam os seus negócios *online*. Isto permite-lhes, para além de criarem uma identidade *online*, tirar partido do crescente número de utilizadores.

Dentro da *web* social encontram-se teias de objetos interligados em sistemas interdependentes. Esses sistemas incluem pessoas, organizações e tecnologias muitas vezes concorrentes (Sapaty, 2019). Do ponto de vista matemático e da ciência da computação, os dados destes sistemas complexos do mundo real podem ser descritos através de redes gráficas formadas por nós interligados (Burguillo, 2018; Savic et al., 2019). Nesse sentido, compreender o comportamento das pessoas, utilizando a *social network analysis* (SNA), para além de desempenhar um papel importante em muitos acontecimentos do mundo real, traz valor acrescentado quer para os académicos, quer para as organizações (Freire et al., 2021).

Apesar dos gestores já utilizarem as RSO de forma rotineira para tomarem decisões melhores e mais rápidas (Laudon e Laudon, 2011) e da sua análise oferecer uma visão cada vez mais detalhada das propriedades relacionais e estruturais da atividade organizacional, a forma de como gerir e analisar a sua aplicação é escassa (Castrucci et al., 2011). Segundo Fu et al. (2017), a investigação existente em RSO ainda não aproveita ao máximo as

oportunidades que estudos interdisciplinares oferecem, porque normalmente se limita às áreas específicas onde é desenvolvida. De acordo com Isson (2018), as empresas são as primeiras a admitir que, apesar de serem ricas em dados, são pobres no conhecimento que deles conseguem extrair.

As empresas demonstram cada vez maior interesse na extração e análise dos dados das RSO para apoio à decisão. Isso deve-se ao facto de os utilizadores deixarem livremente as suas opiniões nas RSO e, de forma (quase) automática e em tempo real, as empresas terem acesso a uma amostra de dados de interações sociais e respetivos conteúdos. Essas interações incluem ligações entre conhecidos, amigos, empresas, etc. e produzem um discurso *web*, em linguagem natural, composto por um conjunto sequencial de trocas discursivas. Assim, as RSO são plataformas onde se podem recolher dados interessantes não só para investigação, mas também para apoio à decisão organizacional.

As empresas devem por isso utilizar as RSO não só para divulgarem e comercializarem produtos, acompanhar e quantificar desempenhos, mas sobretudo para extrair dados para o apoio à decisão. Muitas vezes escondido, o conhecimento que pode ser extraído dos dados das RSO é uma fonte de informação que permite conhecer o consumidor, as suas preferências, tendências, comportamentos, opiniões, etc. A maioria das pessoas concorda que as RSO estão cheias de potencial e têm um efeito positivo nos consumidores (Antunes e Costa, 2012c). Nesse sentido, para as empresas, a análise deste novo tipo de dados permite não só que se diferenciem da concorrência, mas também que criem valor comercial (Isson, 2018).

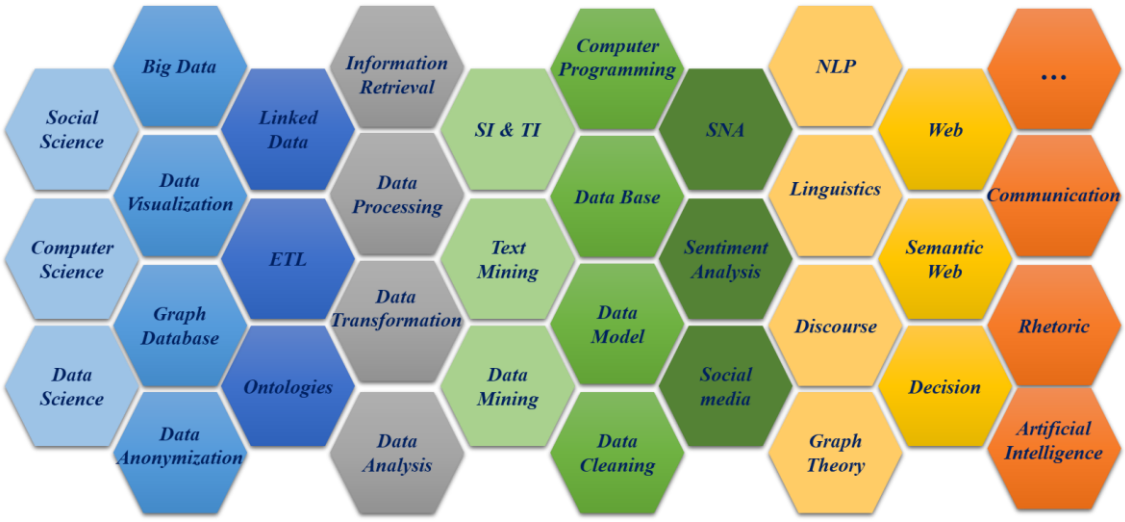
A investigação, análise e aplicação dos dados das RSO, segundo Davenport (2014), tem três grandes contrapartidas para as empresas: redução de custos; melhoria nas decisões e melhoria nos produtos e serviços. Mas, para que as empresas tirem partido destas contrapartidas, devem ser elas próprias a extrair e processar os seus dados sem recorrer a entidades externas, pois, caso contrário, não conseguem perceber o que está a acontecer em tempo real, visto que não têm acesso à informação em tempo útil. Analisar em tempo real, indica a capacidade de processar dados à medida que estes estão disponíveis, em vez de os armazenar e recuperar num qualquer momento do futuro (Barlow, 2013). Para as empresas isto é importante porque dados dinâmicos, em constante mudança, ficam ultrapassados rapidamente se não forem analisados no momento certo.

A extração de informação e conhecimento útil a partir das interações entre utilizadores de RSO é, assim, algo a explorar. Todavia, as trocas discursivas que compõem

o discurso *web* das RSO não são só constituídas por uma estrutura de mensagens. São também constituídas por um sentido construído entre os utilizadores. Desta forma, para analisar o discurso *web* é necessária a interligação de vários domínios de conhecimento tais como o processamento de linguagem natural (PLN)[1], linguística, análise do discurso, *data-mining*, entre outros (Freire et al., 2021).

Estudar as estruturas das RSO remete para a SNA que, de acordo com Moghaddam et al. (2017) é só uma das peças do puzzle, como ilustra a Figura 1. Por seu lado, o discurso *web* remete para a comunicação mediada por computador (CMC) que, de acordo com Herring (2013), se caracteriza por ser uma fusão entre a linguagem escrita e o discurso falado. Para além disso, a interpretação da estrutura do discurso *web*, e o sentido do que é trocado, remete para as teorias clássicas da análise do discurso que, por si só, são áreas multidisciplinares. A análise dos conteúdos textuais remete ainda para o domínio da linguística e, por fim, para a teoria da análise de texto, que pressupõe que a linguagem e o conhecimento podem ser modelados em redes de palavras interligadas, para analisar a presença, frequência e coocorrência de conceitos (Antunes e Costa, 2012a).

Figura 1 - Interdisciplinaridade e multidisciplinaridade da análise de RSO



Conforme já referido, perceber as mensagens faz uso do PLN para que os computadores capturem a estrutura de um texto, interpretem os conteúdos de um documento segundo regras definidas e produzam dados textuais coerentes. Ainda dentro da área das ciências da computação, a extração, armazenamento e processamento dos dados remetem para os SI e as TI a elas associadas. A extração de dados das RSO utiliza tecnologias,

[1] Processamento da linguagem natural (PLN), tradução de *natural language processing* (NLP).

software e linguagens de programação, que auxiliam nesse processo. O armazenamento dos dados das RSO remete para o domínio das bases de dados e o processamento serve-se dos conceitos de *data-mining* e *text-mining*.

Por último, para analisar os dados das RSO torna-se necessário o conhecimento da ciência dos dados. Neste contexto, como referem Moreira et al. (2019), as tecnologias aqui envolvidas permitem extrair conhecimento relevante e útil dos dados obtidos. Um dos objetivos da análise de dados é conseguir, através de técnicas próprias, extrair informação ou conhecimento a partir de um conjunto de dados. Todavia, poucas são as vezes em que o mundo dos negócios disponibiliza dados devidamente estruturados e alinhados para aplicação direta dessas técnicas. A realidade organizacional é muito mais confusa. Ainda assim, apesar de existir alguma complexidade na análise dos dados da *web* social, de acordo com Osei-Bryson e Rayward-Smith (2009), quando os dados são bem explorados e utilizados corretamente podem ser um recurso valioso para a gestão das organizações.

Interligar as várias áreas de conhecimento necessárias para a investigação de RSO no apoio à decisão requer uma perspetiva ampliada, de forma a ultrapassar os limites de uma única área do conhecimento, articulando os saberes dos diferentes domínios de investigação e respetivas interligações. A ausência desta interdisciplinaridade, entre as áreas que analisam as RSO, deve-se ao facto de existirem entre elas diferenças significativas ao nível de foco e em termos teóricos. Isto faz com que diferentes abordagens tenham características interdisciplinares diferentes, pois dependem da influência que sofrem por parte de uma ou outra(s) área(s) do conhecimento. Como referem Fu et al. (2017), o contributo interdisciplinar das ciências sociais, por um lado, e da ciência da computação, por outro, permite avanços significativos na análise das RSO.

Apesar da SNA, metodologia utilizada para a análise das RSO, ter uma longa tradição nas ciências sociais (Alhajj e Rokne, 2018; Borgatti, 2009), a sua interligação com as áreas da computação, do PLN, linguística etc. é uma mais valia recente para produzir conhecimento útil para o apoio à decisão. Como salienta Marmo (2011), a conjugação da metodologia SNA e da mineração *web* dá um grau de detalhe inovador na análise de RSO, que pode ser útil, nomeadamente, ao nível da tomada de decisão. Antunes et al. (2014), por seu turno, consideram que a estruturação do discurso de um grupo permite uma melhor compreensão dos pontos de vista expressos, bem como uma sequência lógica da própria discussão. Contudo, segundo Póvoa et al. (2017), a utilização da SNA para apoio à decisão ainda é uma área de investigação emergente. Também Pozzi et al. (2017) referem que as

RSO representam uma área desafiadora que deve ser analisada para criar conhecimento útil para a tomada de decisão. Nesse sentido, é importante olhar para os conteúdos produzidos pelos utilizadores, nomeadamente para o discurso *web* (Freire et al., 2015b, 2017).

As RSO contêm informação importante sobre as atividades e os conteúdos publicados por utilizadores, que pode ser utilizada tanto em trabalhos académicos, como organizacionais. A utilização dessa informação e publicação dos dados pode levar à divulgação de informação confidencial dos utilizadores das RSO. Todavia, não utilizar dados devido à preocupação com a proteção de dados tornaria qualquer obtenção de informação inútil e a sua análise impossível para apoio à decisão. Por questões éticas e legais, tem de se efetuar a modificação dos dados para que a informação confidencial fique privada, *data anonymization*, (Chester et al., 2018; Raghunathan, 2013; Tripathy e Baktha, 2018) e não correr o risco de uma eventual divulgação ainda que acidental. O objetivo é proteger a privacidade das pessoas e utilizar os dados de acordo com o Regulamento Geral de Proteção de Dados (RGPD[2]).

Só atualmente com a evolução dos SI, das TI e da *internet* e do grande volume de dados de atividade humana, é possível construir e analisar RSO em larga escala (Fu et al., 2017; Savic et al., 2019). Isto pode ser explicado por dois motivos. Em primeiro lugar, porque os custos de armazenamento e processamento de dados têm vindo a diminuir, o *hardware* e o *software* melhoraram e surgiram ferramentas de análise de dados de fácil utilização (Isson, 2018). Em segundo lugar, porque o desenvolvimento tecnológico aumentou as competências em SI e TI do utilizador comum, bem como as dos decisores e gestores, tornando-os mais conhecedores da tecnologia (Freire et al., 2021). Assim, tendo em conta todos os avanços e evoluções tecnológicas dos últimos anos, é possível combinar um conjunto de técnicas, metodologias e *software* para analisar e explorar as RSO de forma mais simples e ágil.

Os trabalhos teóricos que impulsionaram esta investigação apontavam a utilização das tecnologias da *web* semântica como o caminho para interligar as contribuições dos utilizadores das RSO, no apoio à decisão (Antunes e Costa, 2011, 2012b; Antunes et al., 2014). O termo *web* semântica designa um conjunto de tecnologias que contempla

[2] RGPD é o quadro jurídico europeu que entrou em vigor a 25 de maio de 2018 em Portugal, atualmente em execução pela lei nacional n.º 58/2019, de 8 de agosto. Acesso em 22 de novembro de 2019, disponível no portal da Comissão Nacional da Proteção de dados: https://www.cnpd.pt/bin/faqs/faqs.htm.

ontologias[3], agentes[4] de *software* e regras de lógica (Antunes e Costa, 2011). A ideia-chave da *web* semântica é representar os dados e a informação da *web*, utilizando uma linguagem formal passível de ser processada pelas máquinas. No entanto, apesar da *web* semântica ter sido cunhada quase há duas décadas (Berners-Lee et al., 2001), a sua complexa estrutura ainda não responde às necessidades de contextos organizacionais específicos. Segundo Antunes et al. (2016), para responder a estas necessidades seria necessário criar, de forma completa e exaustiva, padrões para todos os conceitos possíveis e existentes na *web*.

Na realidade, ainda não existem padrões que possam ser utilizados de forma generalizada e aplicados aos conteúdos de um contexto específico. As ontologias utilizadas para definir formalmente a estrutura e o significado de dados interpretáveis por máquinas e simultaneamente devolverem, num formato padrão, vocabulário de utilização comum para utilização na *web* semântica, irão preencher essa lacuna. Para o efeito, desde 2004 que o *World Wide Web Consortium* (W3C[5]) assumiu um papel de liderança na expansão da visão inicial da *web* semântica (Riley, 2017), desenvolvendo padrões. Contudo, os padrões do W3C, ainda não são utilizados de forma comum pelas empresas, pois suportam-se em linguagens computacionais complexas tais como o *resource description framework* (RDF[6]), o *SPARQL query language* para *RDF* e a *web ontology language* (OWL).

A crescente importância atribuída pelas empresas à tomada de decisão em contexto *web* requer, por isso, que sejam definidos e implementados mecanismos mais eficientes e mais simples para apoiar as atividades do processo de decisão. Apesar dos atuais avanços tecnológicos bem como das vantagens que os SI e as TI trazem para o apoio à decisão organizacional, o processo de análise dos dados produzidos na *web* social ainda é algo complexo para algumas organizações. Este processo deixa qualquer decisor com um conjunto de dados que tem de transformar, em tempo útil, em informação, para fazer escolhas e/ou estudar opções para tomar decisões melhores e mais rapidamente.

[3] De acordo com a Aufaure et al. (2006), uma ontologia segue um conceito social que é o resultado do acordo sobre a compreensão de um conceito dentro de um domínio de conhecimento por parte de um determinado grupo.

[4] Agentes de *software* são programas que realizam determinadas operações em nome de um utilizador ou de outro programa com algum grau de independência e/ou autonomia e, ao fazê-lo, executam um conjunto de objetivos ou tarefas para as quais foram projetados (Walton, 2007).

[5] O *world wide web consortium* (W3C), é uma comunidade internacional que desenvolve padrões abertos para garantir o crescimento de longo prazo da *web*. Acesso em 14 de agosto de 2013, disponível no portal do W3C: https://www.w3.org/.

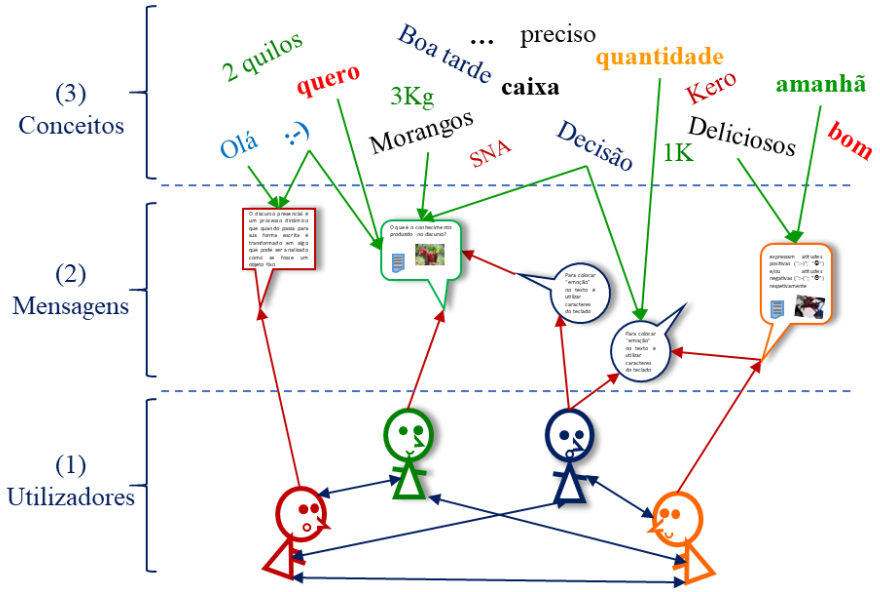
[6] Segundo Walton (2007), a RDF é uma linguagem baseada em *eXtensible markup language* (XML) para representar informação e permite capturar a relação entre um sujeito (recurso) e um objeto (outro recurso).

1.1 Objetivos de investigação

O primeiro objetivo concreto deste trabalho foi desenvolver um modelo para extrair, processar e analisar de forma simples, dados oriundos das RSO, utilizando a SNA, de modo a estruturar a informação do discurso *web* para apoio à decisão. O modelo foi desenvolvido a partir do estudo de caso descrito no subcapítulo 6.2 (TAP). Este estudo de caso permitiu identificar as tecnologias necessárias para estruturação da informação (SI e TI, ligadas à área da computação) e, em simultâneo, validar a aplicabilidade do modelo com dados reais. O modelo forneceu uma base conceptual que considerou não só as várias estruturas de rede incorporadas nos dados das interações sociais e conteúdos semânticos, mas também as técnicas utilizadas para extrair, processar e analisar esses dados. Por esse motivo, este trabalho de investigação abordou a complexidade subjacente aos dados do discurso *web*, focou os aspetos importantes a ter em conta na sua análise, bem como na sua utilização no apoio à decisão no contexto das RSO.

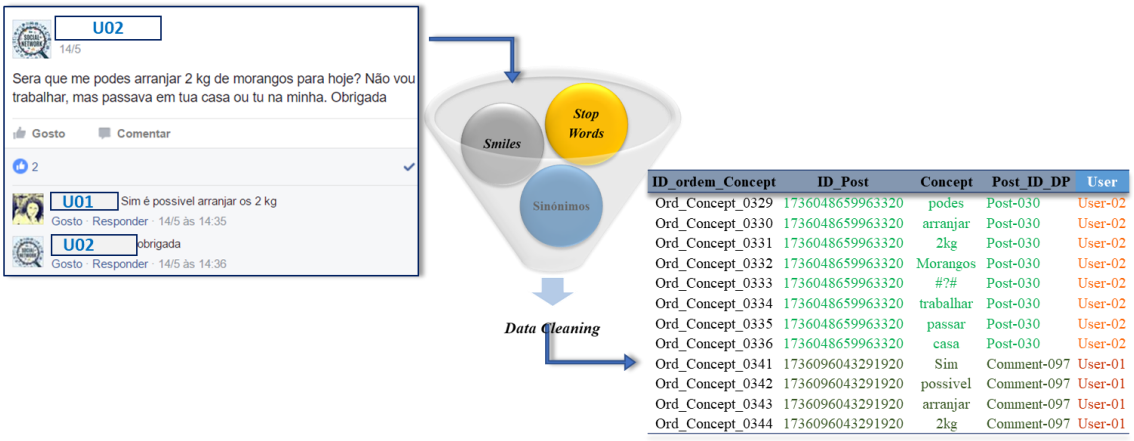
O segundo objetivo foi analisar as interações e o diálogo produzido entre os utilizadores, de forma a extrair informação que auxiliasse no apoio à decisão. A análise das RSO, como representado na Figura 2, caracterizou-se por ter três dimensões ou níveis de análise, cada um com elementos distintos. Os elementos do nível 1 foram definidos como pertencendo à entidade utilizador (*user*), visto que o nível comportava os utilizadores e as interações entre eles. Os do nível 2 foram definidos como pertencendo à entidade mensagem (*post*), pois o nível era composto pelas trocas discursivas entre utilizadores. Finalmente os elementos do nível 3 foram definidos como pertencendo à entidade conceito (*concept*), que se referia ao conteúdo semântico (texto) das trocas discursivas.

Figura 2 - Níveis de análise do discurso web



O terceiro objetivo deste trabalho foi estruturar e analisar os dados semânticos em formato tabular a partir da perspetiva da teoria dos grafos. Cada pedaço de texto contido em cada troca discursiva estruturou-se para que fosse possível produzir dados semânticos. O texto foi transformado em dados interpretáveis pela SNA, partindo cada mensagem em conceitos individuais, conforme exemplificado na Figura 3. Com esta transformação, cada conceito passou a ser um nó da sub-rede semântica que, quando agregada, permitiu construir redes com os três níveis de análise. Os dados semânticos extraídos escondiam informação e, quando interligados com as outras entidades (*user* e *post*) do discurso *web*, permitiram identificar novos padrões nos dados. Tendo como ponto de partida o que os utilizadores escreviam, a análise semântica teve como objetivo encontrar conceitos relevantes entre uma grande diversidade de possibilidades. Por exemplo, quantificar um produto para múltiplas utilizações, tais como preferências, encomendas, vendas, reclamações, etc. Obter dados semânticos a partir dos conteúdos produzidos pelos utilizadores foi uma tarefa desafiadora devido a vários problemas: as características da linguagem que cada um utiliza; o contexto onde são produzidas as mensagens; e as dificuldades inerentes não só à recolha, mas também ao processamento dos dados. Para além disso, os dados recolhidos podem ser incompletos, imprecisos ou completamente inadequados à tarefa de processamento semântico para posterior utilização no apoio à decisão. A resolução destes problemas passou por recolher a maior quantidade possível de dados para, em seguida, lhes aplicar técnicas mais ligadas às ciências da computação tais como PLN, *data-mining*, *text-mining* e técnicas de inferência para conseguir informação útil e fidedigna.

Figura 3 - Estruturação do discurso *web*



ID_ordem_Concept	ID_Post	Concept	Post_ID_DP	User
Ord_Concept_0329	1736048659963320	podes	Post-030	User-02
Ord_Concept_0330	1736048659963320	arranjar	Post-030	User-02
Ord_Concept_0331	1736048659963320	2kg	Post-030	User-02
Ord_Concept_0332	1736048659963320	Morangos	Post-030	User-02
Ord_Concept_0333	1736048659963320	#?#	Post-030	User-02
Ord_Concept_0334	1736048659963320	trabalhar	Post-030	User-02
Ord_Concept_0335	1736048659963320	passar	Post-030	User-02
Ord_Concept_0336	1736048659963320	casa	Post-030	User-02
Ord_Concept_0341	1736096043291920	Sim	Comment-097	User-01
Ord_Concept_0342	1736096043291920	possivel	Comment-097	User-01
Ord_Concept_0343	1736096043291920	arranjar	Comment-097	User-01
Ord_Concept_0344	1736096043291920	2kg	Comment-097	User-01

O quarto objetivo foi obter dados estruturados em formato de grafo, quer das interações entre utilizadores, quer de dados semânticos. Este resultado teve duas finalidades: (i) a criação/obtenção de padrões para o processamento semântico, porque para uniformizar os conceitos foi necessário comparar os dados extraídos com um padrão predefinido; (ii) a criação/obtenção de dados estruturados, quer da interação entre utilizadores, quer semânticos, expressivos e representativos de um contexto organizacional para apoio à decisão.

Sabendo, como apresentado anteriormente, que a análise das RSO é multidisciplinar e a forma de tratar os seus dados encontra-se fragmentada pelas ciências sociais e computacionais, o quinto e último objetivo foi evidenciar a interdisciplinaridade entre as áreas de investigação. O modelo proposto pretendeu simplificar e estruturar todo o processo de análise de dados oriundos das RSO. Este processo é considerado, por alguns autores, como sendo difícil de realizar com os métodos atuais de tratamento de dados, nomeadamente a análise semântica do discurso *web*. Como referem Smith et al. (2009a), extrair sentido do grande volume de dados das organizações através da SNA é um grande desafio. A realização deste objetivo permitiu criar uma linha orientadora para investigações futuras, com as áreas de investigação associadas ao conjunto de tarefas que devem ser seguidas na análise de dados das RSO para apoio à decisão. A utilização do modelo permitiu garantir a coerência e integridade de execução, mas também a extrair dados mais fiáveis e credíveis dos conteúdos da *web* social.

1.2 Contribuições

A principal contribuição deste trabalho foi o desenvolvimento de um modelo para transformar dados das RSO em informação útil para apoio à decisão, com ênfase nos conteúdos semânticos. Para validar o modelo e mostrar que se alcançaram os objetivos estabelecidos, realizaram-se quatro estudos de caso. Dois dos estudos de caso, subcapítulos 6.1 e 6.3, utilizaram resultados quantitativos, obtidos com as métricas da SNA, para decisões definidas *a priori*. Nos dois outros estudos de caso e dos subcapítulos 6.2 e 6.4, os dados foram explorados no sentido de perceber a insatisfação dos clientes e a identificação de conflitos entre os clientes e a organização.

Contrariamente a outros analisados na literatura, o modelo desenvolvido (Freire et al., 2017), incorporou os três níveis e as três entidades de análise do discurso *web* num único ambiente. O modelo incluiu não só as características relacionais do utilizador, mas também as técnicas e métodos de transformação e análise semântica de dados de um contexto organizacional específico, onde os dados sociais são não-estruturados e informais. Assim, a aplicação do modelo permitiu executar uma análise das RSO, agregando a estrutura de interação entre utilizadores e aproveitando os conteúdos semânticos de tais interações.

Mais especificamente, o contributo da agregação no modelo dos três níveis demonstrou-se através de cinco aspetos: (i) apreender a estrutura social inerente ao contexto em análise que é fundamental para um estudo abrangente das RSO e dos seus conteúdos; (ii) oferecer às organizações a oportunidade de integrar facilmente os dados sociais nas várias fases do processo de decisão; (iii) ser aplicável quer em contextos organizacionais de pequena quer de grande dimensão; (iv) identificar quem interage com quem (relação *user* | *user*), quem iniciou uma troca discursiva, quem comentou e quem disse o quê; (v) para além da visualização dos dados numa única rede, permitir várias granularidades de análise, ou seja, desagregar os dados e representar várias sub-redes.

Relativamente às questões técnicas e métodos de transformação e análise de dados, o contributo apresenta-se por cinco aspetos: (i) todas as fases do modelo executam-se com ferramentas (*software*/plataformas) sem custos e facilmente manipuladas e utilizadas em contexto organizacional, permitindo garantir uma análise *end-to-end*[7], onde todas as etapas do processo de análise de dados são realizadas com sucesso (extração, processamento e

[7] *End-to-end* descreve um processo que executa todas as etapas do início até ao fim, sem necessidade de obter nada de terceiros, e devolve um resultado funcional e definido (Tickoo e Iyer, 2017).

operações de análise de dados); (ii) a estrutura do modelo simplifica a complexidade da extração, processamento e análise de dados, para obter resultados para apoio à decisão, fornecendo um método sistemático que permite produzir e analisar várias redes possíveis, quer de utilizadores quer semânticas; (iii) a análise da contribuição dos dados semânticos (conceitos) tem duas finalidades: auxiliar a alimentação de uma base de dados de limpeza (*cleaning database*[8]) e, por sua vez, a criação de informação útil para apoio à decisão; (iv) o modelo incorpora uma *graph database*[9] que organiza os dados e permite extrair e criar possíveis sub-redes, de acordo com o objetivo da análise, porque cada entidade corresponde a uma sub-rede e a quantidade de nós e tipos de relacionamento de cada uma é diferente; (v) quando comparada a estrutura do modelo com outras que integram dados sociais, o poder do modelo demonstra-se pela sua aplicabilidade tanto a grandes como a pequenos volumes de dados.

Este trabalho de investigação também contribui para diminuir a fragmentação existente na análise das RSO para apoio à decisão. O desenvolvimento do modelo enquadrou os problemas de investigação de forma transversal, utilizando abordagens e conhecimentos teóricos distintos das áreas das ciências sociais e da computação, para que os dados representem corretamente a informação de um contexto organizacional específico.

Os trabalhos desenvolvidos e apresentados (Antunes et al., 2014, 2016, 2018; Freire et al., 2015a, 2015b, 2015c, 2017, 2021) ao longo desta investigação, contribuíram também para as áreas de conhecimento que interligam as RSO e o apoio à decisão. Os estudos de caso realizados, com dados reais, mostraram a aplicabilidade do modelo desenvolvido, em contextos organizacionais específicos.

1.3 Relevância e justificação do trabalho de investigação

Atualmente é reconhecido que a informação contida nas RSO tem cada vez mais valor e que esse facto, por si só, justifica a investigação e a criação de modelos mais adequados para extração, processamento e análise dos dados da *web* social. Para Sloan e Quan-Haase (2017), o grande volume e variedade de dados que as RSO disponibilizam faz com que a sua investigação seja relevante para muitas disciplinas. A própria literatura

[8] A c*leaning database* é uma estrutura que guarda listas de conceitos definidos como padrão, para limpar e/ou corrigir dados, melhorando assim a sua qualidade.

[9] As *Graph database* gerem dados ligados entre si onde cada registo representa o relacionamento entre dois nós Robinson et al. (2015).

salienta e justifica a pertinência e a necessidade de mais investigação que interligue as RSO e o apoio à decisão, quer atendendo à importância dos temas sobre RSO e o apoio à decisão, quer atendendo à magnitude e ao impacto que podem ter para as organizações. Nesse sentido, são vários os motivos que tornam relevante e que justificam esta investigação.

Em primeiro lugar, a investigação que interliga as RSO ao apoio à decisão é sobretudo teórica e, por isso, a literatura a ela associada mostra contribuições ainda fragmentadas pelas áreas das ciências sociais e da computação. Embora segundo vários autores (Alhajj e Rokne, 2018; Fu et al., 2017; Missaoui et al., 2017) se trate de uma área multidisciplinar, falta uma interligação mais clara entre os vários âmbitos. Associada a essa fragmentação está, por um lado, a diversidade de problemas de investigação que podem ser estudados e, por outro, sendo uma área de investigação cujo tema é abordado por académicos de domínios diferentes, está fundamentada através de teorias, métodos e perspetivas diferentes.

A evolução da *web*, segundo Ozyer et al. (2019), tem influenciado a investigação em RSO para apoio à decisão, fazendo com que esteja menos assente nas ciências sociais e se suporte mais nas abordagens multidisciplinares associadas às ciências da computação. Não existe escassez de literatura sobre RSO, análise dos seus diferentes tipos de topologia, dos seus utilizadores, da sua aplicabilidade e relevância. Contudo, apesar da abundância de investigação nesta área e consequente literatura associada, ainda há conhecimento insuficiente de como lidar com os dados das RSO para que, de forma ágil, suportem os processos de apoio à tomada de decisão.

A fragmentação existente na literatura revela, por isso, que existe uma lacuna na investigação para um entendimento mais profundo dos processos de extração, processamento e análise de dados para apoio à decisão nas suas várias fases. Nesse sentido, continua a ser oportuna a investigação suportada por teorias e metodologias mais ligadas às ciências da computação que, de forma simples, articule e estabeleça uma relação entre os processos de apoio à decisão e os dados das RSO. Como se lida com dados considerados complexos (Barlow, 2013; Pozzi et al., 2017; Roy et al., 2014), as áreas das ciências da computação auxiliam todo o processo de transformação dos dados iniciais, em informação e conhecimento para apoio à decisão.

Como segunda justificação para esta investigação, há também a falta de evidências de como as interações entre utilizadores e os seus conteúdos semânticos, como um todo, podem ser modelados e operacionalizados para apoio à decisão. Na perspetiva de Golbeck

(2015), os conteúdos (texto, foto, vídeos etc.) são as "peças" mais importantes numa investigação profunda no âmbito das RSO.

Medir e quantificar o número de interações entre utilizadores e a quantidade de mensagens que trocam é simples, existe já muita investigação para o efeito na área das ciências sociais. Todavia, a dificuldade está na extração de informação e conhecimento dos conteúdos produzidos pelos utilizadores, visto que os dados não se encontram estruturados à partida. A própria literatura atual (Isson, 2018; Moreira et al., 2019; Shum et al., 2011; Xhafa et al., 2015) salienta que ainda existe uma grande dificuldade na estruturação deste tipo de dados. O conhecimento teórico necessário para executar estas tarefas é do domínio das ciências da computação.

Assim, a investigação em RSO tem como ponto fraco o processamento semântico dos conteúdos textuais. Isso prende-se com o facto de o estilo de escrita utilizado nas RSO ter vulgarmente um padrão fora do comum. Ou seja, os utilizadores cometem com frequência, erros ortográficos e/ou gramaticais, utilizam abreviaturas, símbolos, incluem *links*, imagens, áudio, vídeo, etc. Para além disso, como referem Jokinen e Wilcock (2017), as novas tecnologias afetam a utilização da linguagem e esta, por seu lado, adapta-se às mudanças que a tecnologia provoca. O processamento de dados textuais apresenta por isso desafios de vários níveis, tais como lidar com a codificação de caracteres, até inferir significado de um contexto específico (Ingersoll et al., 2013; Isson, 2018).

A diversidade dos conteúdos das RSO está associada um potencial de informação para apoio à decisão que é possível extrair, o que, como terceiro motivo, também justifica a pertinência desta investigação. Em particular, os dados produzidos nas RSO integram uma estrutura emergente que é inerente às ligações entre utilizadores, às suas interações aos grupos a que pertencem. O *Facebook*, o *Twitter*, o *Instagram*, o *WhatsApp* entre outras, são exemplos de RSO da *web* social onde todos os utilizadores (pessoalmente ou representando as suas organizações) têm igual poder quando adicionam ou modificam conteúdos. Apesar destas plataformas diferirem na sua forma de colaboração, elas têm em comum uma estrutura social onde as contribuições coletivas (conteúdos) representam as crenças e/ou convicções que os utilizadores partilham entre si. Davenport (2014) defende que provavelmente os dados mais relevantes para grandes decisões são os dados da *web* social, pois contêm o que as pessoas ao redor do mundo andam a dizer e a fazer, o que é importante para elas, com quem interagem e o porquê (Golbeck, 2015). Assim sendo, nas plataformas de RSO surgem padrões de contribuição e agregação de informação semelhantes. Nesse sentido, também é

importante a criação de modelos que não se limitem exclusivamente a apurar a posição a favor ou contra uma ideia, mas que agreguem de forma inteligente as contribuições individuais para apoio à decisão.

O quarto motivo que justifica esta investigação é o facto da maioria das ferramentas para apoio à decisão, no âmbito das RSO, se apoiar e depender de metodologias e tecnologias complexas. Se, por um lado, assentam em estruturas complexas, tais como sistemas de apoio à decisão (SAD), modelos de argumentação, *web* semântica, ontologias, bases de dados (entre outras), por outro, requerem mais recursos tecnológicos, humanos e financeiros. Nesse contexto, as metodologias e tecnologias utilizadas, devido a sua complexidade, não permitem interligar e aceder em tempo útil aos utilizadores, aos seus dados e redes a que pertencem, para os medir ou monitorizar de forma eficaz.

O quinto motivo, associa o volume de dados e a diversidade de fontes como pertinentes de investigação. Como refere a literatura, devido à crescente popularidade das RSO, produzem-se dados a um ritmo alucinante (Missaoui et al., 2017; Moreira et al., 2019; Savic et al., 2019; Tannen, 2013; Xhafa et al., 2015). Para além disso, estão em diferentes formatos e fontes (Bahga e Madisetti, 2019; Stieglitz et al., 2018; Tiroshi et al., 2017). De acordo com Roy et al. (2014), explorar dados complexos de diferentes fontes e extrair o seu sentido é um problema muito desafiador.

O aumento do volume de dados, bem como a sua diversidade, implica que não seja simples, com técnicas manuais, agregar, processar e analisar os dados para apoio à decisão, nem tomar decisões rápidas e eficientes. Campbell et al. (2013) referem que o volume e natureza não estruturada dos conteúdos das RSO apresentam vários desafios, mas que são uma fonte rica de dados para análise. Também Samanthula e Jiang (2014), reforçam que é importante perceber como se devem explorar os dados das RSO no sentido de recuperar informação útil para o apoio à decisão organizacional. Segundo Robinson et al. (2015), como os métodos de processamento da informação continuam a evoluir, a próxima barreira a ultrapassar reside na capacidade de capturar, analisar e compreender as interligações dos dados.

A sexta e última questão, que justifica a relevância da investigação, está relacionada com a forma como as teorias e modelos tradicionais de apoio à decisão olham para os seus processos e fases. Os modelos clássicos de tomada de decisão estão associados à existência de processos racionais e sequenciais em contextos organizacionais tradicionais. As teorias e modelos de decisão tradicionais assentam na ideia de que, perante um acontecimento, existe

um problema e é necessário identificar alternativas para o resolver. No entanto, a informação extraída de dados das RSO para apoio à decisão em tempo real está constantemente a mudar, melhorar e a evoluir. Este facto faz com que atualmente, na era dos dados, se inverta o paradigma. Isto é, na sequência de um acontecimento o(s) problema(s) podem não estar à partida identificado(s) e explorar os dados permite não só identificá-lo(s), mas também antecipá-lo(s) ou simplesmente ver tendências de mercado ou identificar perfis de utilizadores para direcionar produtos ou serviços.

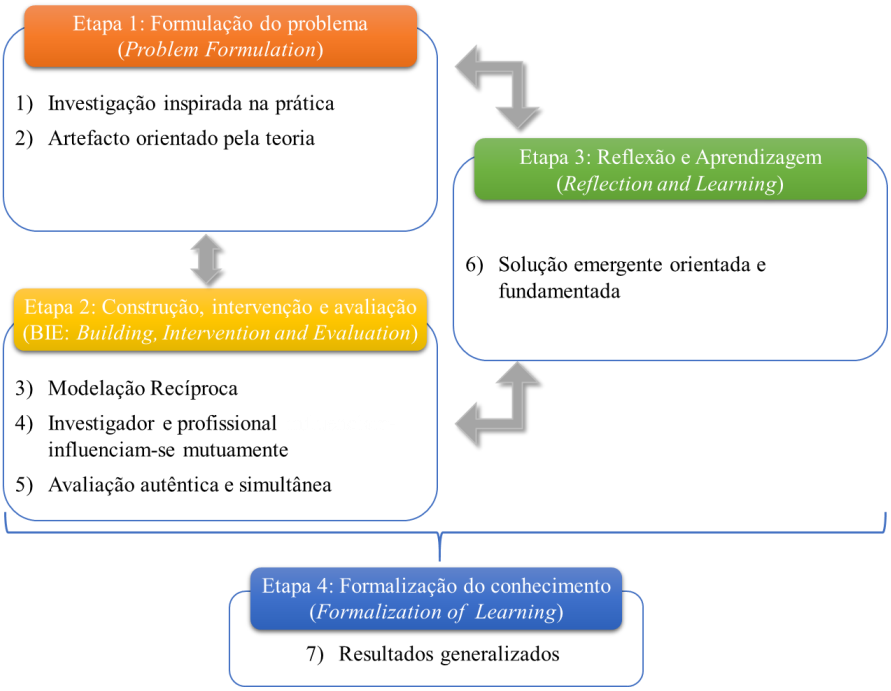
O processo de decisão assente em dados das RSO pretende ser mais preditivo e analítico, diferente do processo de decisão tradicional onde só *a posteriori* se atuava. O objetivo, como referem Moreira et al. (2019), não é prever o que vai acontecer no futuro, mas sim o quão prováveis são os resultados de um determinado acontecimento. Também segundo Bahga e Madisetti (2019), a análise preditiva, para além de perceber a frequência ou resultado de um acontecimento, tem como finalidade responder à questão: "O que é provável que aconteça?". Nesse sentido, o conhecimento e valor extraído dos dados das RSO através da SNA, pode fornecer um novo discernimento e algumas oportunidades que auxiliem as várias fases dos processos de tomada de decisão.

1.4 Metodologia

O caráter multidisciplinar da análise de dados das RSO para apoio à decisão remete não só para as ciências sociais, mas também para as ciências da computação. Com o objetivo de obter dados estruturados como solução para problemas práticos, considerou-se que a metodologia *action design research* (Sein et al., 2011) era a mais apropriada para abordar os problemas e objetivos desta investigação.

Do ponto de vista teórico, a *action design research* permitiu interligar o conhecimento de vários domínios das ciências sociais e computacionais, que deram à investigação em RSO fundamentação metodológica para estruturar as etapas de extração, armazenamento e processamento de dados. Devido à complexidade que uma investigação desta natureza pode atingir, seguiu-se a *action design research* (Sein et al., 2011) que ofereceu uma estrutura de atividades suportadas por vários princípios, apresentados na Figura 4, que se executaram ao longo do trabalho de investigação. Assim, a aplicação da *action design research* foi essencial, para este trabalho que teve como objetivo não só desenvolver um modelo conceptual, mas também produzir informação para apoio à decisão.

Figura 4 - Etapas e princípios da *action design research*



Fonte: adaptado de Sein et al. (2011)

O que justificou a escolha da *action design research* foi o facto do seu pragmatismo oferecer benefícios para a resolução de problemas organizacionais reais, pois é apropriada para a investigação de problemas de natureza prática ligados aos SI (Hevner et al., 2004; Mullarkey e Hevner, 2018; Sein et al., 2011). Para além disso, esse pragmatismo considera que as consequências práticas ou as consequências reais são elementos essenciais para o apuramento quer do significado, quer da verdade (Hevner, 2007).

Para alguns autores (Haj-Bolouri et al., 2018; Johannesson e Perjons, 2014; Luftenegger et al., 2015), a *action design research* é uma versão da *design science research*[10] de Hevner et al. (2004) que, para além de articular a *design science*[11] e a *design research*[12], privilegia as influências do contexto organizacional na criação e na evolução de um artefacto. Outros autores (Mettler, 2015; Mullarkey e Hevner, 2018) defendem que a *action design*

[10] Por definição a *design science research* é uma atividade de investigação que constrói artefactos novos e inovadores, para resolver problemas ou melhorá-los, ou seja, cria meios para alcançar um objetivo geral não necessariamente especificado (Hevner et al., 2004).

[11] A *design science* enfatiza que uma investigação deve ser motivada por um problema prático, abordado por uma metodologia rigorosa, uma avaliação de acordo com padrões científicos, e os resultados devem ser comunicados à comunidade científica (Hevner et al., 2004).

[12] A *design research* tem como objetivo desenvolver uma solução funcional para aplicação num problema de natureza prático e os resultados são relevantes apenas para um dominio específico, na qual apenas um indivíduo, grupo ou organização estão envolvidos (Hevner e Chatterjee, 2010).

research é uma mistura da *action research*[13] e da *design science research* pois as duas são semelhantes, se não idênticas, sugerindo que, em algumas situações e contextos, as duas podem ser articuladas. Para Luftenegger et al. (2015), a *action design research* requer uma estreita colaboração entre o contexto académico e o organizacional.

Independentemente da forma como é utilizada a *action design research*, numa fase inicial foi necessário posicioná-la dentro do contexto prático do trabalho para investigar as questões a resolver, assim como identificar as melhores estratégias e métodos que deveriam ser utilizados. Foram também efetuadas comparações com outras metodologias existentes na literatura, no sentido de melhorar e evoluir os conceitos e as suas aplicações. A metodologia *action design research* utilizou-se como instrumento para a construção de um artefacto, que visou resolver problemas de natureza prática em contexto real e, desta forma, levar à criação de conhecimento.

No âmbito deste trabalho de investigação, à medida que se identificavam e recolhiam as necessidades organizacionais de cada contexto de análise, e do(s) problema(s) de investigação, a *action design research* serviu de base para a construção e desenvolvimento de um artefacto. Para além disso, serviu também para o seu aperfeiçoamento e evolução. Durante o processo foi necessário avaliar o artefacto, bem como justificar e fundamentar a sua importância. Estas atividades, construção, desenvolvimento, avaliação e justificação, do ponto de vista teórico, suportaram-se na base de conhecimento académico já existente. Assim, a base de conhecimento inicial estabeleceu-se através da fundamentação e de métodos já consolidados e reconhecidos pela literatura dos diferentes domínios teóricos (Hevner et al., 2004; Sein et al., 2011).

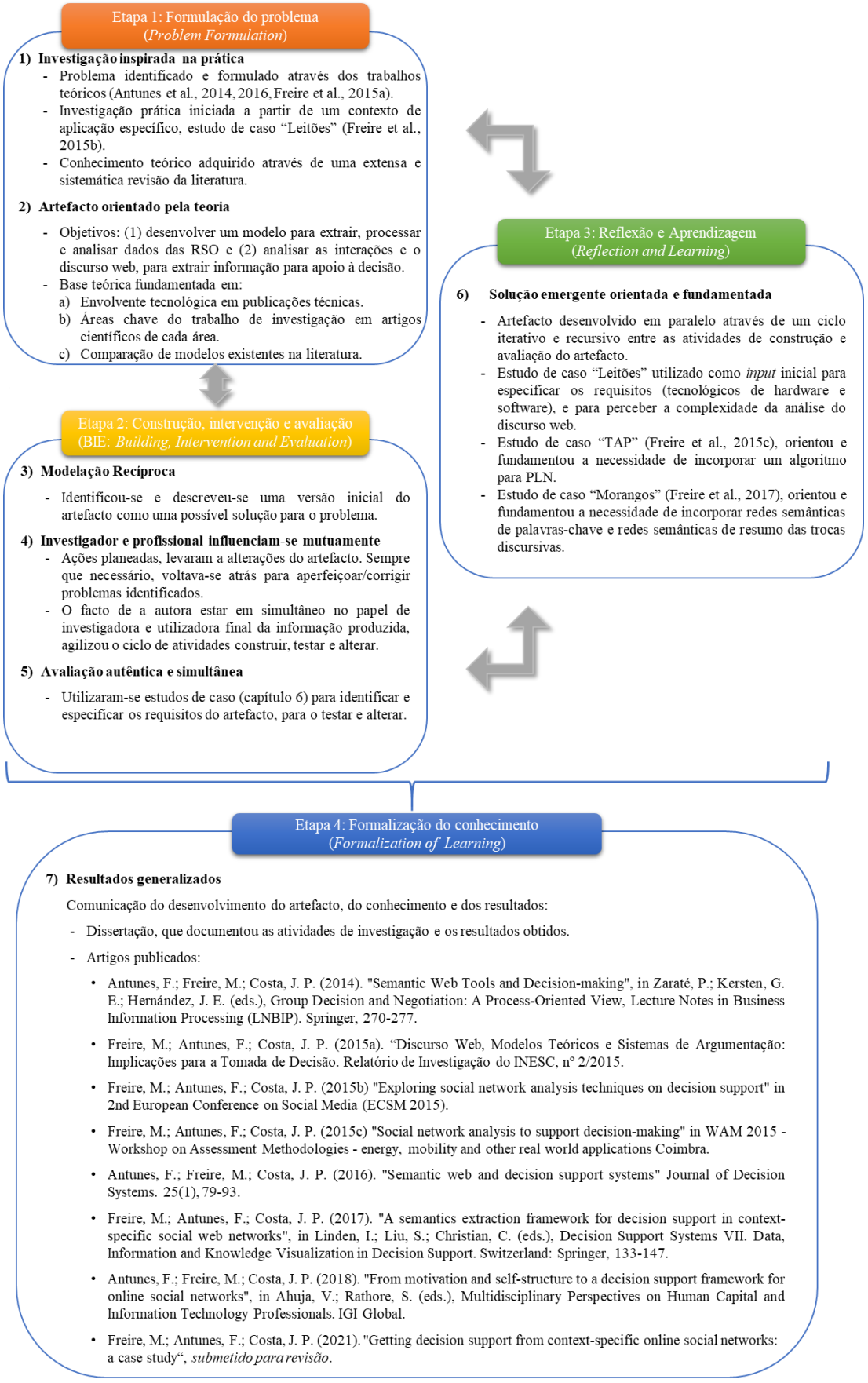
Desta forma, a *action design research* foi vista como um tipo de *design research* (investigação por projeto) que privilegia as influências de cada contexto organizacional na construção e na evolução do artefacto e seguiu um ciclo de aperfeiçoamento através da recursividade sistemática entre o agir e o investigar. Esse ciclo, composto pelas etapas construção, intervenção e avaliação (Sein et al., 2011), teve como objetivo um constante aperfeiçoamento tanto da aplicação prática do modelo proposto como da própria investigação. Como a *action design research* inclui igualmente a criação de conhecimento fundamentado, dentro de contextos específicos e práticos, enquadra-se também no âmbito

[13] A *action research* tem o duplo objetivo de contribuir tanto para a prática como para a teoria e de assumir a existência e o envolvimento de um cliente concreto Livari e Venable (2009).

da *design science*. Para além disso, visto que a *action design research* pressupõe uma participação ampla das partes (teórica e prática), houve uma interação onde o conhecimento local (específico de cada contexto) não só se misturou com o geral, mas também com o especializado e técnico.

Dentro da panóplia de técnicas de *design* existentes na literatura, considerou-se a *action design research*, Sein et al. (2011) como a mais adequada para abordar os problemas inerentes à análise de RSO para apoio à decisão. Na Figura 5 resumem-se as etapas e princípios que foram considerados na realização do presente trabalho.

Figura 5 - Etapas e princípios do processo de investigação do presente trabalho



1.4.1 Primeira etapa: formulação do problema

Sein et al. (2011) explicam que nesta fase deve-se identificar e compreender o problema de investigação para o qual se pretende apresentar uma solução. Assim, nesta etapa o objetivo foi estabelecer uma investigação inspirada na prática (princípio 1) e um artefacto orientado pela teoria (princípio 2).

Após a identificação do problema através dos trabalhos teóricos (Antunes et al., 2014, 2016; Freire et al., 2015a), iniciou-se a investigação num contexto de aplicação específico. Para o efeito, criou-se o estudo de caso "Leitões" descrito no subcapítulo 6.1 (Freire et al., 2015b). As especificidades e caraterísticas desse estudo de caso não serviram apenas de *input* para especificar os requisitos da investigação, mas também levaram a uma visão inicial da complexidade associada a todo o processo de análise de dados das RSO para apoio à decisão, quer em termos teóricos quer práticos. O capítulo 1 deste trabalho também sustenta esta etapa da *action design research* visto que apresenta o problema, fundamenta teoricamente a necessidade de uma solução para o mesmo e apresenta os estudos de caso a partir dos quais foi emergindo o modelo proposto.

O conhecimento teórico necessário para consolidar as áreas de conhecimento que interligam as RSO ao apoio à decisão obteve-se através de uma extensa revisão da literatura. O objetivo foi adquirir uma base teórica suficientemente fundamentada e estruturada, que suportasse a execução e o desenvolvimento das diferentes fases do modelo proposto. A base teórica fundamentou-se com publicações técnicas que sustentaram a envolvente tecnológica de cada área de conhecimento utilizada, artigos científicos com modelos para comparação e fundamentação da necessidade do modelo proposto.

O caráter multidisciplinar da análise de dados para apoio à decisão envolveu o estudo de uma diversidade de domínios científicos tanto das ciências sociais como computacionais (decisão, linguística, *data-mining*, *text-mining*, bases de dados, entre outras). A revisão de literatura permitiu criar uma base de conhecimento teórico, científico e uma seleção de métodos que alicerçaram esta investigação, garantindo não só o seu rigor, mas também a sua inovação (Hevner, 2007).

1.4.2 Segunda etapa: construção, intervenção e avaliação

Esta etapa da *action design research* teve como objetivo identificar e descrever uma versão inicial do artefacto como uma possível solução para o problema explicado e definido anteriormente.

O princípio (4), que suporta esta etapa, parte do pressuposto que o investigador e o profissional se influenciam mutuamente. De acordo com Sein et al. (2011) pressupõe uma participação não apenas dos investigadores, mas também das partes interessadas em torno de uma ação planeada, na forma de uma intervenção com mudanças da situação do que se está a investigar. O facto de a autora estar em simultâneo no papel de investigadora, profissional e utilizadora final da informação produzida, permitiu agilizar o ciclo de atividades de construir, testar e alterar. Para além disso permitiu atingir um duplo objetivo. O primeiro, desenvolver o modelo apresentado no capítulo 5. O segundo, obter dados fiáveis para apoio à decisão. Exemplo disso são os resultados obtidos em cada estudo de caso. Esses objetivos atingiram-se não só articulando as sinergias e as contribuições mútuas entre a pertinência da utilidade prática da solução e o seu rigor científico, mas também resolvendo os problemas de natureza prática que iam surgindo. Como refere Hevner (2007), a utilidade prática por si só não define uma boa investigação científica.

Para efeitos de avaliação do artefacto, os recursos teóricos considerados foram soluções semelhantes propostas ou implementadas por outros estudos académicos (descritos no subcapítulo 4.5.4), em conjunto com os requisitos que abordavam, assim como os interesses e objetivos deste trabalho de investigação. Do ponto de vista prático, os estudos de caso foram a estratégia de investigação utilizada para não só identificar e especificar os requisitos iniciais, mas também para testar e alterar em conformidade o modelo proposto.

De acordo com a literatura, a avaliação define-se como um processo exigente de teste e verificação do comportamento do artefacto no ambiente para o qual foi idealizado e projetado. Nesta etapa, um dos objetivos foi avaliar o artefacto, modelo proposto, e a solução que o mesmo oferecia para a resolução do(s) problema(s) prático(s) apresentado(s) neste trabalho, juntamente com o nível de cumprimento dos requisitos identificados em cada estudo de caso.

Para testar o modelo proposto em situações reais de apoio à decisão, realizaram-se mais três estudos de caso (subcapítulos 6.2, 6.3 e 6.4) que permitiram o aperfeiçoamento do modelo e a sua aplicabilidade em diferentes cenários. Isso permitiu não só desenvolver, mas

também aplicar o artefacto em dados reais de contextos organizacionais reais e identificar o impacto do modelo.

Utilizaram-se os estudos de caso visto que este tipo de estratégia tem o seu foco num acontecimento específico e oferece uma descrição e uma visão rica e profunda do mesmo (Johannesson e Perjons, 2014; Yin, 2018). Assim, os estudos de caso utilizaram-se para obter uma melhor compreensão das práticas e restrições técnicas e tecnológicas, bem como do contexto específico de cada caso e dos recursos analíticos necessários e associados à complexidade dos problemas e das necessidades da autora enquanto investigadora.

Uma das atividades fundamentais em qualquer investigação empírica é a recolha de dados acerca do fenómeno ou do acontecimento sob investigação (Johannesson e Perjons, 2014). Na estratégia de extração de dados utilizada, encontra-se implícita as técnicas de recolha de dados digitais produzido nas RSO. Do ponto de vista da elaboração do modelo, esta fase de recolha de dados teve como objetivo obter dados digitais de *input* adequados e utilizáveis nas restantes fases do modelo proposto. Do ponto de vista da aplicabilidade do modelo em casos reais, esta fase teve como objetivo obter dados para os transformar em informação útil para apoio à decisão.

A estratégia de análise de dados teve como objetivo extrair informação importante dos dados no sentido de descrever ou explicar as situações dos estudos de caso sob investigação (Johannesson e Perjons, 2014). Os dados originais por si só não falam e, antes de ser possível tirar conclusões, um dos requisitos para que os mesmos possam ser utilizados foi a sua preparação, interpretação e análise. Assim, para resolver problemas de complexidade computacional e interoperabilidade, aplicou-se aos dados um pré-processamento recorrendo a técnicas do *data-mining* e *text-mining*, para estruturar os dados de *input* do modelo. A SNA auxiliou na análise quantitativa de dados com recurso às métricas de centralidade e na análise qualitativa através da visualização gráfica, ambas interligadas aos estudos de caso e muito importantes na análise de dados (Yin, 2018).

1.4.3 Terceira etapa: reflexão e aprendizagem

Esta terceira etapa da *action design research*, princípio (6) solução emergente orientada e fundamentada, desenvolveu-se, quase em paralelo (senão em paralelo), a partir da interação entre a etapa um e dois à medida que iam emergindo as versões do artefacto, isto é, do modelo proposto. Assim, o artefacto foi sendo desenvolvido através de um ciclo

que era o resultado da interação entre as atividades de construção e avaliação tanto do artefacto como dos seus processos Hevner (2007).

O modelo proposto como artefacto para RSO no apoio à decisão encontra-se descrito no capítulo 5. O desenvolvimento do artefacto seguiu as melhores práticas e as técnicas das áreas necessárias para o desenvolvimento de cada fase do modelo proposto. Para além disso, articularam-se os requisitos tecnológicos da ciência da computação e análise de dados com os conceitos das ciências sociais para criar o artefacto desejado e necessário. A articulação das várias estratégias e métodos, estudos de caso, extração de dados e a análise de dados, contribuíram para uma visão mais abrangente e uma descrição mais completa dos fenómenos sociais (Hollstein, 2014).

1.4.4 Quarta etapa: formalização do conhecimento

Um dos requisitos e objetivos de âmbito genérico da *action design research* é que o artefacto desenvolvido seja apropriado e relevante não só para um contexto em particular, mas sim para vários. A implementação do modelo proposto, nos estudos de caso, demonstrou que o mesmo pode ser utilizado em diferentes contextos, envolvendo diferentes situações de apoio à decisão. Exemplos disso foram a sua aplicação num simples evento departamental, no contexto do transporte aério, venda de morangos e comércio automóvel.

Para a prova de implementação do modelo, utilizaram-se três tipos de dados (estruturados, semiestruturados, não estruturados) e múltiplas fontes de dados (GDF, CSV, XLS). Não houve restrições quanto ao tipo de decisão de cada estudo de caso que o modelo pôde manipular, para além da restrição de alimentar a *cleaning database* com dados de cada contexto organizacional. Essa restrição, garantiu que a análise semântica era a apropriada para a transformação dos dados semiestruturados ou não estruturados (texto).

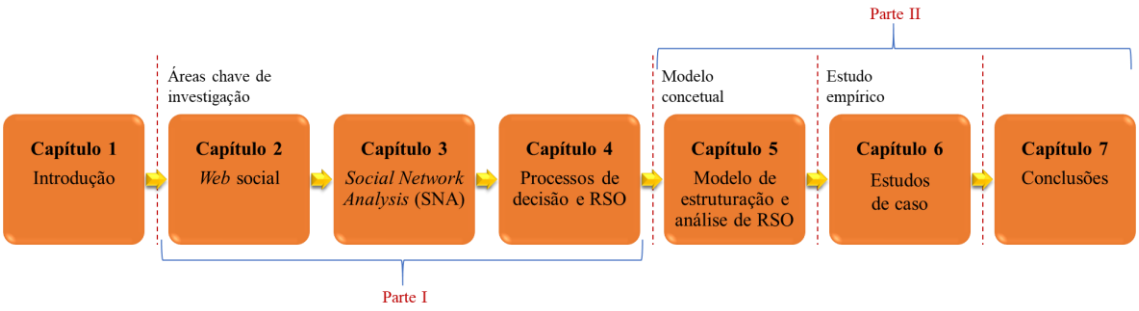
A última etapa de todo o processo de investigação, comunicação do artefacto, visa divulgar, não só à comunidade científica, mas também à prática, o conhecimento e os resultados produzidos referentes ao artefacto proposto (Hevner, 2007; Sein et al., 2011). Nesse âmbito, enquadram-se os artigos publicados apresentados em Antunes et al. (2014, 2016, 2018); Freire et al. (2015a, 2015b, 2015c, 2017, 2021) assim como esta dissertação, que teve o objetivo de documentar as atividades de investigação e os resultados obtidos durante o percurso da mesma. Os estudos de caso realizados, com dados reais, mostraram a

aplicabilidade do modelo desenvolvido (artefacto), em contextos organizacionais práticos específicos.

1.5 Estrutura do documento

O presente trabalho de investigação encontra-se dividido em sete capítulos. Após a introdução, onde se expôs a relevância, os objetivos da investigação e as suas contribuições, bem como a metodologia utilizada, seguem-se os capítulos 2, 3 e 4 que abordam os temas chave da investigação. O capítulo 5 apresenta o modelo de análise das RSO, no capítulo 6 descrevem-se os estudos de caso realizados no âmbito deste trabalho de investigação e no capítulo 7 apresentam-se as conclusões finais. A organização do trabalho está representada na Figura 6.

Figura 6 - Estrutura do trabalho de investigação



No capítulo 2 abordou-se o discurso *web*, produzido nas plataformas das RSO no seio da *web* social. Devido à complexidade associada à análise dos conteúdos produzidos no discurso *web*, este foi enquadrado nas teorias clássicas da análise do discurso, para explicar os fenómenos discursivos da comunicação (Freire et al., 2015a). Assim, estabelecendo um paralelismo com o discurso *web*, os utilizadores assumem a figura de oradores e/ou interlocutores, as mensagens ou trocas discursivas definiram-se como atos de fala e os conteúdos como recursos linguísticos e não linguísticos. O objetivo foi utilizar estes elementos para traduzir de forma formal o ambiente *online*. As várias abordagens da análise do discurso ajudaram a explicitar formalmente a estrutura discursiva, quer ao nível da interação entre os utilizadores, quer ao nível dos conteúdos que produzem, podendo representar-se em unidades de análise de modo a facilitar a sua codificação e criação de conjuntos de dados semânticos.

No capítulo 3 discutiram-se os conceitos chaves da metodologia SNA e apresentaram-se as métricas utilizadas para analisar e representar as RSO. Para além disso, caracterizaram-se os dados das RSO, assim como os *softwares* existentes para extração e análise de dados bem como de visualização das RSO. Apresentou-se também uma visão geral de estudos de análise e modelos originados em trabalhos de investigação no âmbito das RSO, concluindo-se que os estudos que utilizam a SNA e a análise semântica dos *posts* são poucos e sobretudo teóricos, pelo que a investigação prática ainda está muito concentrada em artigos e livros, que abrangem apenas os aspetos metodológicos e matemáticos da análise das RSO. Para além disso, como se entra no domínio da linguística, do discurso, do PLN, entre outras áreas, a maioria dos estudos acaba por ter o seu foco na descrição das propriedades da rede de utilizadores e não integra redes semânticas. Estudos mais complexos aprofundam a análise semântica através da lente da *web* semântica, mas não interligam as redes semânticas com a rede de utilizadores. Foram também analisados trabalhos de investigação com modelos para análise das RSO, no sentido de identificar diferenças entre eles e o modelo proposto em Freire et al. (2017).

No capítulo 4 tratou-se a utilização das RSO para apoio à decisão, nomeadamente através da recolha direta das opiniões e experiências de um maior número de pessoas sem serem necessários inquéritos formais. A sua análise, para efeitos de apoio à decisão organizacional, tem de considerar métodos e ferramentas quer numa perspetiva organizacional, quer tecnológica. Em que a perspetiva organizacional está relacionada com os processos de decisão e a tecnológica com as RSO com o ambiente tecnológico.

No capítulo 5 descreveu-se o modelo proposto em Freire et al. (2017, 2021), de análise de dados das RSO para apoio à decisão. O modelo tem em conta que, com os avanços e evoluções tecnológicas que têm acontecido nos últimos anos, é possível utilizar e combinar um conjunto de técnicas, metodologias e *software* para analisar e explorar os dados das RSO, de forma mais ágil e sem recurso a tecnologias muito complexas. Inicia-se com a extração de dados, seguindo-se um processo recorrente e iterativo que produz dados que permitem representar várias redes e sub-redes. O modelo encontra-se dividido em três fases sequenciais relacionadas entre si. Na primeira fase extraem-se os dados das interações entre utilizadores (*user* | *user*), ligações entre utilizadores e mensagens (*user* | *post*) e as mensagens e comentários que trocam. Na segunda fase, armazenamento e transformação dos dados, utilizam-se métricas da SNA para interpretar e eliminar dados irrelevantes. Recorre-se à visualização dos dados para analisar sistematicamente os resultados obtidos e verificar se

são necessários ajustes. Na terceira fase, utiliza-se o *text-mining* e a SNA para efetuar a análise semântica e extrair e resumir informação útil, à partida não estruturada. Os *outputs* de dados finais têm dois objetivos: (i) alimentar uma base de dados de limpeza (*cleaning database*), com conceitos padronizados, para processar os dados não estruturados; (ii) traduzir a informação para apoio à tomada de decisão de forma compacta e visualmente "atraente".

A ideia que está subjacente neste modelo é a de minimizar e atenuar as atuais desvantagens da *web* semântica, desenvolver e aplicar uma abordagem que permita extrair e tratar dados não estruturados (texto). O modelo utiliza dados com formatos diferentes, recolhe-os e estrutura-os para que seja possível responder às questões e aos problemas de decisão de um contexto organizacional específico.

No capítulo 6 apresentaram-se os estudos de caso analisados ao longo deste trabalho. O objetivo destes casos foi perceber como estruturar dados das RSO, utilizando a metodologia SNA, de modo a extrair informação do discurso *web* produzido pelos seus utilizadores para o apoio à decisão. Para o efeito recolheram-se dados de dois grupos criados e geridos pela autora e dados das *FanPages*[14] de duas grandes empresas (Transportes Aéreos Portugueses (TAP), fabricante de veículos Volkswagen). Nos grupos criados conheciam-se os utilizadores, o contexto de decisão e a linguagem utilizada. Nas *FanPages* desconhecia-se, à partida, que tipo de dados se iriam encontrar e que informação iriam mostrar. Os estudos de caso utilizaram-se com dois objetivos: a investigação do apoio à decisão orientada por dados das RSO e a identificação das áreas de conhecimento necessárias para elaboração de um modelo de análise das RSO. De seguida resumem-se os quatro estudos de caso: (i) estudo de caso "Leitões" (Antunes et al., 2018; Freire et al., 2015b); (ii) estudo de caso "TAP" (Antunes et al., 2018; Freire et al., 2015c); (iii) estudo de caso "Morangos" (Antunes et al., 2018; Freire et al., 2017, 2021) e (iv) estudo de caso "Volkswagen".

A anonimização é cada vez mais importante na gestão e análise de dados à medida que os dados das RSO são continuamente extraídos, a proteção da privacidade dos utilizadores tem de ser considerada na execução dos processos de *data-mining*.

No trabalho aqui apresentado, todos os dados extraídos das RSO eram de acesso público, livremente disponibilizados pelos utilizadores em grupos fechados, criados em

[14] *FanPages*: páginas específicas do Facebook de empresas ou marcas (Villegas, 2016).

páginas pessoais, ou por empresas em páginas institucionais (*FanPage*). Nos dados dos quatro estudos de caso, por questões éticas e legais, as designações dos utilizadores foram anonimizadas. De acordo com a literatura (Chakraborty e Tripathy, 2016; Tripathy e Baktha, 2018) existem várias técnicas de anonimização. No entanto podem ser divididas em duas categorias: mascarar os dados e decompor em tabelas. Para o estudos de caso "Leitões", o procedimento de anonimização executou-se com recurso à primeira técnica, referida por Raghunathan (2013), isto é, substituíram-se os nomes dos utilizadores por outros caracteres. Para os restantes estudos de caso, a anonimização implementou-se de acordo com a técnica de decomposição em tabela referida em Nassar et al. (2015). Nesta técnica, os dados considerados sensíveis separam-se, na base de dados, dos restantes e armazenam-se em tabelas diferentes. Para cada estudo de caso, os dados confidenciais foram estruturados em formato tabular e foi atribuído um *ID* sequencial, a cada utilizador que foi identificado. No processamento dos dados utilizaram-se apenas os *IDs* e os critérios relacionais definidos entre a tabela de utilizadores e as restantes.

O estudo de caso "Leitões", Freire et al. (2015b) e Antunes et al. (2018), teve como objetivo perceber a complexidade subjacente à análise do discurso *web*, os aspetos importantes a ter em conta na sua análise bem como a sua utilização no apoio à decisão no contexto das RSO. Para o efeito utilizou-se uma pequena rede de utilizadores para permitir limitar a quantidade de elementos em análise.

No sentido de obter dados reais para análise, criou-se um grupo fechado na rede social *Facebook* e colocou-se em discussão os detalhes para um almoço social após uma reunião de empresa. O grupo era composto por colaboradores de um departamento de uma empresa nacional portuguesa, geograficamente dispersa por todo o país. Os colaboradores aderiram voluntariamente ao grupo com diferentes níveis de participação. A organização de um simples almoço tem associado um processo logístico que envolve, a escolha da localização, condicionada pelo número de pessoas que estará presente, bem como as preferências para o tipo de comida e espaço de forma a satisfazer os participantes. Por esse motivo, a tomada de decisão exigiu uma pesquisa minuciosa das alternativas sugeridas e informações sobre as preferências, para obter os resultados pretendidos. O facto de se limitar o objeto de estudo e de se estar inserido nesse objeto, significou o conheciamento da linguagem utilizada, da semântica associada e respetivos conceitos não sendo necessárias ontologias para os "traduzir" e sendo fácil decifrar o que está implícito no discurso.

No estudo de caso "TAP", Freire et al. (2015c) e Antunes et al. (2018), o foco da análise dos dados foi a insatisfação do cliente e a identificação de problemas durante um período de dez dias de greve (maio de 2015). Para este tipo de problema, por norma, as empresas utilizam pequenas amostras de dados e recorrem a inquéritos para medirem e analisarem a insatisfação em momentos de crise. Os objetivos deste estudo foram os de analisar as interações e conteúdos produzidos por clientes, que eram utilizadores externos à companhia aérea e que por isso não partilhavam um sentimento de grupo e/ou objetivos comuns. O objetivo principal foi perceber se era possível estruturar as mensagens, produzidas na *FanPage*, a fim de verificar o serviço e a capacidade de resposta da companhia aérea, bem como desenvolver indicadores para rever e/ou reforçar as estratégias de atendimento ao cliente em situações de greve, utilizando a SNA. Os dados recolhidos, não apenas durante os dez dias de greve, mas também nos dois dias seguintes, permitiram construir várias redes e sub-redes. Através da análise do conteúdo semântico das mensagens percebeu-se que os mesmos podiam ser tipificados por categorias, o que significa que forneciam informação para direcionar/criar grupos de atividades para tratar assuntos específicos tais como reclamações, solicitações de ajuda e/ou apoio na alteração de voos e pedidos de informações (sobre hotel, recolha de bagagem, procedimentos de reembolso, etc.). Nas mensagens analisadas, muitos clientes reclamavam que o *call center* não atendia as chamadas, não recebiam resposta aos *e-mails*, existiam balcões de atendimento presencial fechados nos aeroportos e a página *web* da companhia aérea não dispunha de informação atualizada sobre a greve. A visualização das redes semânticas permitiu identificar conversas relevantes e obter informação em tempo útil da opinião dos clientes, o que permitiria atuar mais rapidamente no sentido de não deixar agravar o sentimento de insatisfação dos clientes. A análise das mensagens e a visualização das redes (quer de utilizadores, quer semânticas) poderiam ter ajudado a companhia aérea a identificar problemas de comunicação nos canais tradicionais (*call center, e-mail* etc.) e a resolvê-los mais prontamente, no sentido de evitar futuras ruturas de serviço por motivo de greve. Para além disso, a visualização e a identificação de palavras-chave poderiam ter sido utilizadas para reforçar a atualização da informação nos canais *online* e/ou balcões de atendimento no que se refere: à realização de listas de voos cancelados e/ou disponíveis; procedimentos para reagendar a data dos voos dos clientes.

No estudo de caso "Morangos", Freire et al. (2017, 2021) e Antunes et al. (2018), o problema de apoio à decisão era estimar a quantidade necessária de morangos para

satisfazer as encomendas a entregar a clientes, exclusivamente com base nas suas mensagens do *Facebook*. Pretendeu verificar-se se a análise semântica das mensagens podia ser preditiva no sentido de recomendar as quantidades necessárias a transportar e a entregar. À medida que os potenciais clientes faziam encomendas, em paralelo contabilizaram-se e registaram-se manualmente as quantidades encomendadas para posterior validação dos dados. Concluiu-se que era possível extrair informação para apoio à decisão, ou seja, antecipar o comportamento dos clientes com base em dados reais, decorrentes da interação e trocas discursivas entre os utilizadores. Através das métricas da SNA não foi apenas possível fazer *ranking* de quem fazia mais pedidos, identificou-se também a estrutura da rede de utilizadores, bem como a sua representação gráfica. As informações obtidas ajudaram a criar estratégias de marketing direcionadas, envio de mensagens, para utilizadores-chave que imediatamente as partilhavam. Os efeitos foram um aumento dos pedidos sem um custo adicional na divulgação do produto. Concluiu-se que não só era possível recomendar as quantidades necessárias a serem transportadas e a entregar, mas também identificar os clientes que mais encomendavam e os líderes do grupo de clientes.

No estudo de caso "Volkswagen" (VW), o foco da análise dos dados foi perceber o impacto do escândalo de emissões de poluentes da VW (em setembro 2015) na RSO *Facebook*. Na sequência do escândalo, a VW despediu quadros de topo por presumível envolvimento na instalação de *software* que adulterava as medições de consumo e de emissões poluentes nos veículos. A situação fez disparar os custos do grupo por vários motivos: multas, reparações, retomas, etc. Atualmente, opiniões e informações negativas sobre as marcas disseminam-se muito rapidamente nas RSO e o caso da VW não foi exceção. Neste estudo, utilizaram-se várias métricas e visualizações da SNA para ilustrar o impacto deste caso. O objetivo principal foi estruturar as mensagens, produzidas na *FanPage* da VW, para perceber a reação dos utilizadores, clientes ou não da marca, bem como desenvolver indicadores para monitorizar as conversas *online*. Este tipo de "escuta" das RSO permite que as marcas identifiquem informação útil, com o objetivo de limitar danos de imagem e evitar que um acontecimento evolua para uma crise e/ou desastre. Permite também que as empresas estejam sempre informadas sobre o que as pessoas dizem sobre elas, e onde, permitindo assim ações direcionadas em situações críticas. O caso aconteceu a 18 de setembro de 2015 e os dados foram recolhidos em períodos temporais de uma semana, entre 31 de agosto de 2015 e 31 de janeiro de 2016. Os mesmos permitiram construir várias redes e, com isso, descobrir os temas mais discutidos logo após o conhecimento do caso.

A análise dos conteúdos semânticos e a identificação de palavras-chave (tais como: *emissions*, *problem*, *affected*) permitiu obter informação relevante dos tópicos em discussão e perceber que quem entrava na *FanPage* da VW à procura de uma qualquer outra informação sobre a marca ficava inevitavelmente exposto à história do caso.

Este trabalho de investigação termina no capítulo 7 onde se apresentam as conclusões e as indicações para investigações futuras. Referem-se ainda as limitações encontradas na execução deste trabalho, visto que foi partir delas que se identificaram oportunidades que podem vir a ser consideradas futuramente.

Por último, refere-se que no material suplementar[15] a este trabalho, encontram-se as figuras e as tabelas em formato JPG, bem como extratexto de suporte a este trabalho.

[15] Disponível em: https://meocloud.pt/link/c927c457-f6a2-47ab-bbad-5def596a5bd8/Tese-Phd-CAD/.

2 Web social

A interação social entre grupos de pessoas é uma componente chave da vida humana que estava limitada às restrições de tempo e espaço, antes do aparecimento da *Web*. Essas restrições foram parcialmente eliminadas com o aparecimento e com a evolução da *Web*, bem como com a sua disseminação. A *Web*, como atualmente a conhecemos, evoluiu de uma versão inicial (denominada *Web* 1.0), na qual os utilizadores simplesmente atuavam como editores ou consumidores de conteúdos, para a *Web* 2.0 onde todos os utilizadores são, em simultâneo, editores e consumidores de informação (Antunes et al., 2014).

De acordo com a literatura, a versão da *Web* designada por *Web* 2.0 é considerada a base inicial quando pensamos numa versão participativa e colaborativa da *Web*, onde os utilizadores são capazes de se envolver e criar conteúdos, desenvolvendo uma inteligência coletiva (Beer, 2009; Lytras et al., 2009; Shimazu e Shinichi, 2007). Como referem Alhajj e Rokne (2018), na *Web* 2.0, comparativamente à *Web* 1.0, os utilizadores deixaram de ser consumidores passivos da informação e tornaram-se autónomos e interativos, ou seja, produtores, editores e distribuidores de informação. Este facto levou a um aumento do volume de dados, justificado na literatura pelo aumento do volume dos conteúdos criados pelos utilizadores, que são assim uma das características que definem a *Web* 2.0.

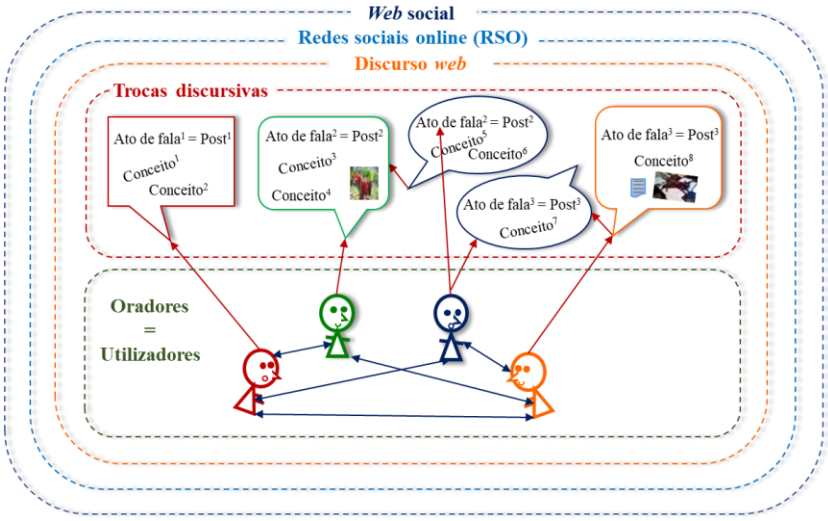
No seio da *Web* 2.0 existe uma área definida como *web* social que agrega as plataformas das RSO e que permite às pessoas comunicarem umas com as outras. Jamali e Abolhassani (2006) consideram que a *web* social, por si só, também pode ser considerada uma rede social, visto que é formada por páginas e *links* para outras páginas da própria *web*. No entanto, Turban et al. (2018) fazem a diferenciação entre os dois conceitos, definindo o conceito de *web* social como o conjunto de ferramentas e tecnologias e incluindo no conceito de rede social as pessoas interligadas (utilizadores), as suas interações, o suporte social e o conteúdo digital por elas criado.

Assim, os recursos tecnológicos oferecidos pela *web* social têm vindo a proporcionar novas perspetivas e ferramentas para a comunicação *online*. A topologia de rede descentralizada e de baixa hierarquia, das RSO, permite novas possibilidades quer de consumo, quer de divulgação de conteúdos. Em particular, o aparecimento de ferramentas e tecnologias *Web* e a sua evolução para a *Web* 2.0 permitiram que as pessoas se organizassem em RSO, da mesma forma que se organizam em redes sociais no mundo real. Segundo

Laudon e Laudon (2011), os gestores utilizam de forma rotineira as chamadas tecnologias *Web* 2.0, tais como as RSO, ferramentas de colaboração e *wikis*, a fim de tomarem melhores decisões e também de forma mais rápida.

Como representado na Figura 7, a *web* social é composta por três dimensões. De uma forma mais lata, a primeira dimensão é a própria *web* social que envolve as plataformas e ferramentas da *Web* 2.0. A segunda dimensão, integra as plataformas de redes interpessoais que permitem a comunicação e interação social, isto é, as designadas RSO. A terceira dimensão é o discurso *web*, através do qual os utilizadores trocam e partilham mensagens, ou seja, produzem informação. Assim, as RSO reúnem utilizadores que partilham ideias, permitindo a criação de comunidades em torno dos mesmos interesses. Para além da sua dimensão social, as RSO funcionam também como um facilitador da aprendizagem formal e informal.

Figura 7 - Dimensões da *web* social



2.1 Discurso *web*

Atualmente, a conjugação da linguagem humana e da tecnologia tem criado novas formas e utilizações da linguagem (Jokinen e Wilcock, 2017). O discurso produzido nas RSO não é apenas uma forma de comunicação e interação social *online*, mas é, essencialmente, uma fonte de informação.

De acordo com Jorgensen e Phillips (2002), Van Dijk (1985) e Wooffitt (2005), a diversidade de métodos descritivos de análise do discurso é impressionante e a linguagem pode ser analisada através de várias perspetivas, onde cada área olha de forma diferente para

este discurso. Assim, no contexto das RSO, a análise do discurso *web*, para além de envolver a gramática (ou seja, regras para falar corretamente uma língua), também envolve as tecnologias que auxiliam o seu processamento e interpretação formal. Todavia, para formalizar em termos computacionais os conteúdos produzidos pelos utilizadores são necessários os conhecimentos teóricos das ciências sociais ligadas ao discurso. Para analisar o discurso *web* são necessários métodos de análise do discurso das teorias tradicionais, conjugados com áreas mais recentes das ciências da computação, tais como o PLN, os SI e as TI, entre outras.

No que se refere ao próprio conceito de discurso, é possível encontrar na literatura ligada às ciências sociais uma grande diversidade de definições. Em particular, Jaworski e Coupland (2006) apresentam dez definições recolhidas de um conjunto amplo de fontes. Todas essas definições recaem em três categorias principais: (i) o discurso é qualquer coisa para além de uma frase, (ii) o discurso é a utilização da linguagem e (iii) o discurso é um amplo conjunto de práticas sociais que inclui fatores não linguísticos.

No contexto das RSO, o discurso caracteriza-se por ser uma fusão entre a linguagem escrita e o discurso falado. Nesse sentido, Herring (2011) define o discurso *web* como sendo a comunicação produzida entre os seres humanos que interagem uns com os outros, transmitindo mensagens, utilizando computadores ligados em rede. Herring (2013) sugere que o discurso produzido no âmbito da *Web* 2.0 se poderia chamar de *convergent media computer – mediated discourse* (CMCMD) ou "Discurso 2.0". A autora justifica a designação quer pela existência de novos tipos de conteúdos que devem ser analisados, quer pela diversidade de novos contextos que também devem ser considerados.

No âmbito da *web* social, o discurso *web* é definido como uma especialização dentro da CMC, visto que utiliza os seus recursos digitais. Bodomo (2010) define a CMC como sendo a codificação e a descodificação de símbolos linguísticos entre emissor e recetor, com o objetivo de processar a informação existente em vários meios de comunicação *online* (*internet*, *e-mail*, mensagens de texto, *YouTube*, *Skype*, entre outros). Todavia, para efeitos de análise semântica e interpretação da linguagem natural, o discurso tem de ser definido tanto de uma forma mais restrita, onde se incluem os níveis sintáticos e semânticos, como numa mais ampla, onde se incluem o contexto e as interações entre as pessoas. Para além destas duas formas, Perelman e Olbrechts-Tyteca (2005) e Wooffitt (2005) defendem que a análise do discurso também deve considerar a retórica.

Relativamente à forma restrita, as definições clássicas do discurso podem aplicar-se para a análise semântica do conteúdo das mensagens do discurso *web* pois, como refere a literatura (Van Dijk, 1977; Van Dijk, 1985), na sua forma restrita, a análise do discurso é um constructo abstrato, que investiga a linguagem falada ou escrita pela ótica da linguística e só tem em conta a estrutura da mesma (ao nível sintático e semântico). Para a construção das redes semânticas do modelo proposto no capítulo 5, foi necessário ter em consideração a sintaxe e a semântica para definir regras formais que automaticamente interpretassem o conteúdo das mensagens. A sintaxe foi necessária na construção de redes semânticas de resumos para dividir as mensagens em conceitos individuais e definir a sequência em que aparecem no texto. A semântica foi necessária para classificar formalmente os conceitos, isto é, definir padrões. Estes definiram-se de acordo com seu contexto organizacional específico, ou seja, a semântica própria de cada um. Se por um lado interessava a forma como as partes do texto se relacionavam entre si para construir uma unidade de significado, por outro, interessavam os processos linguísticos, cognitivos e sociais através dos quais os significados eram expressos e as intenções da interação humana interpretadas (Jaworski e Coupland, 2006).

No que se refere à forma ampla, as definições clássicas do discurso podem aplicar-se para a análise das interações entre os utilizadores do discurso *web* e para caracterizar o contexto organizacional onde ocorrem. Para analisar as interações entre os utilizadores foi necessário entender como os processos de produção, aceitação e compreensão do discurso *web* ocorriam. A produção está relacionada com a quantidade de conteúdos publicados, a aceitação e compreensão com a partilha de conteúdos ou com um simples *like*[16]. Tudo isto é quantificável utilizando as métricas da SNA. Para além disso, se em cada contexto organizacional existe um tipo específico de discurso, por outro, cada discurso é característico de um contexto específico. Isto foi importante para a análise do discurso *web*, visto que perceber cada contexto organizacional permitiu ajustar especificamente o significado dos conceitos e interpretar as interações entre utilizadores. Permitiu, também, analisar o conhecimento que estava a ser produzido e que ia circulando no decorrer do discurso *web*. Segundo Gee (2001), este conhecimento desencadeia visualizações mentais, pressupostos e padrões de conversação, que têm a ver com as crenças e ações que compõem as práticas sociais dentro de um determinado contexto.

[16] O botão *like* foi implementado pelo *Facebook* para permitir que os utilizadores interagissem com uma publicação e apareceu em 2009, sendo depois replicado para todas as RSO (Russell e Klassen, 2019).

A retórica, enquanto subcampo da análise do discurso, tem o seu foco na utilização de estruturas especiais de texto e conversação tais como as metáforas[17], as comparações, a ironia, etc., ou seja, o tipo de estruturas designadas de figuras de estilo. Assim, a retórica foi importante para a análise do discurso *web*, visto que o discurso utiliza artifícios de linguagem como estratégias discursivas para, através da persuasão, captar a atenção da audiência. Um artifício simples de linguagem, característico do discurso *web*, é a utilização de *smiles*[18] como forma de concordância, corroboração ou apreciação do que os outros partilham. Ao contrário de outras estruturas de texto e conversação, estas estruturas retóricas são facultativas no discurso e utilizam-se para transmitir ou produzir efeitos específicos, como parte de uma estratégia de persuasão. Estes artifícios de linguagem podem enfatizar significados e, assim, chamar a atenção dos destinatários do discurso *web*. Como refere Wooffitt (2005), a retórica preocupa-se menos com a forma como a linguagem é utilizada, ou seja, não tem o seu foco na análise da conversação e na análise do discurso, preocupando-se muito mais com o caráter argumentativo e/ou persuasivo do discurso. A retórica é a capacidade de utilizar, em cada situação e de forma discreta, um artifício de linguagem convincente, para influenciar os outros intervenientes do discurso *web* e com isso controlar o fluxo de informação.

O termo análise do discurso utiliza-se para descrever uma diversidade de abordagens de investigação que têm o seu foco no uso da linguagem (Lomax, 2004). Existem diferentes tipos de análise do discurso, tais como a análise da conversação, a análise crítica do discurso e a análise de conteúdos (Wooffitt, 2005). Cada uma delas tem os seus próprios métodos e premissas, mas o ponto em comum é a análise e construção do significado do discurso enquanto forma de comunicação. Todas elas foram importantes, visto que a análise da semântica contida no discurso *web* pode, de um modo geral, definir-se como o estudo da linguagem enquanto meio de comunicação da *web* social.

Relativamente à análise da conversação, para Gardner (2004) e Wooffitt (2005), esta difere das outras abordagens no que concerne a alguns pressupostos teóricos, princípios metodológicos e técnicas de análise. De acordo com os autores, a análise da conversação tem o seu foco na conversação, em particular nos atos de fala e não na linguagem, dando

[17] A metáfora é uma figura retórica que só é reconhecida no seu contexto, pois a sua estrutura não é nem gramatical nem semântica, mas deve-se a uma relação com alguma coisa que não é o objeto imediato do discurso (Perelman e Olbrechts-Tyteca, 2005).

[18] *Smiles*, na CMC é a prática de utilização de ícones para transmitir emoções e outras características linguísticas (Bodomo, 2010).

ênfase à interação social entre os intervenientes. Nesse sentido, definiram-se os atos de fala como sendo as mensagens trocadas pelos utilizadores nas RSO. Formalmente, podem ser referidas como o vínculo entre um ato de fala e o utilizador que o criou. Como refere Van Dijk (1977), os atos de fala do discurso são o que cada interveniente diz alternadamente e estabelecem-se através da interação entre eles. Em senso comum a ideia é "agora falo eu, depois falas tu". De igual modo, para Jurafsky e Martin (2009), o discurso caracteriza-se pela troca de atos de fala, onde primeiro fala o interveniente A, depois o B, em seguida novamente o A e assim por diante. Estes atos de fala podem ser observados em unidades e formalmente codificados para posterior análise, seja de um único conceito, de uma frase ou de um conjunto de frases onde a função locutória é ocupada pelos diversos utilizadores.

De acordo com Gardner (2004), a análise da conversação envolve alguns elementos fundamentais. O primeiro elemento é a noção de interação, pois o discurso (conversação) é visto como uma atividade conjunta entre o ouvinte e o orador, onde ambos estão em pé de igualdade enquanto construtores do diálogo que vai emergindo. No discurso *web*, o utilizador (enquanto orador) projeta as suas contribuições diretamente para os destinatários (também eles utilizadores) e, por sua vez, estes podem influenciá-lo com as mensagens que devolvem. Nesta atividade, cada troca discursiva (ou unidade de conversação) é entendida pelos participantes, tendo em conta a compreensão que têm do que foi dito anteriormente. Na análise do discurso *web*, as trocas discursivas codificaram-se formalmente como unidades de conversação e definiram-se como sendo a ligação entre uma troca discursiva e a resposta dada à mesma.

O segundo elemento, referido por Gardner (2004), é a noção de tempo entre atos de fala, ou seja, o silêncio e/ou o falar simultaneamente durante uma conversa. Na análise do discurso *web*, um silêncio (uma mensagem sem resposta) pode afetar a codificação e interpretação dos dados recolhidos, porque os dados dizem respeito a um período temporal definido. Pode acontecer que quando uma mensagem é comentada passados vários dias por outro(s) utilizador(s), os dados recolhidos não incluam esses comentários. Em contrapartida, as trocas discursivas consecutivas podem indicar que os outros estão a dar relevo ao que estava a ser dito.

A consideração do tempo abre questões relativas quer à forma como as trocas discursivas estão organizadas entre os participantes, quer à forma de codificação em cada período temporal. Quando aplicada ao discurso *web*, a análise da conversação teve como objetivo descrever como as várias mensagens se combinavam entre si. Isto permitiu obter

uma visão não só da forma como os utilizadores construíam linguisticamente as mensagens (no que se refere à sintaxe e semântica), mas também da forma da estrutura da sua sequência, para codificação e interpretação. Wooffitt (2005) reforça esta ideia quando refere que a análise da conversação é um método que permite reproduzir, de forma coerente, as interações que ocorrem na *web* social.

No que se refere à análise crítica do discurso, Van Dijk (2001) e Wooffitt (2005) indicam que esta se preocupa com o modo pelo qual as estruturas do discurso influenciam as representações mentais, visto ser uma análise que tem o foco no poder que o discurso exerce sobre os outros. Para Herring (2011, 2013), a análise crítica do discurso é um dos métodos utilizados para analisar o discurso *web* ao nível dos fenómenos sociais. Nas RSO, o discurso *web* de alguns utilizadores tem a capacidade de formar e influenciar opiniões. Esta capacidade, na *web* social, pressupõe que quem tem controlo sobre o fluxo de informação é importante, tem proximidade e influência sobre os outros e tem prestígio (*status*, popularidade). No âmbito do discurso *web*, a análise crítica do discurso de forma explícita permitiu analisar e expor os utilizadores-chave existentes dentro da RSO. Para Van Dijk (2001), num contexto específico, certos significados e formas de discurso têm mais influência sobre a mente das pessoas do que outros, como sugere a própria noção de persuasão da teoria retórica. Segundo Locke (2004), é possível influenciar as pessoas, com a informação que veem como sendo a mais importante num texto ou discurso e assim chegar aos seus modelos mentais. Segundo Wodak e Meyer (2009), a análise crítica do discurso não está interessada na análise das unidades linguísticas em si, mas em estudar os fenómenos sociais de poder e persuasão que, pela sua complexidade, requerem uma abordagem multidisciplinar.

Outra área importante para a análise do discurso *web* foi a análise de conteúdo. Segundo Van Dijk (1985), esta pressupõe a utilização de categorias para análise qualitativa, tais como palavras, frases e características estilísticas, podendo envolver outras unidades de análise para identificar a frequência com que aparecem, através de métodos estatísticos. De acordo com Rubin e Babbie (2011), a análise de conteúdo é uma forma de transformar dados qualitativos em dados quantitativos, que pode ser aplicada a praticamente qualquer tipo de comunicação e é composta principalmente pela codificação das ocorrências dos conteúdos que estão a ser comunicados. A análise do conteúdo foi considerada como um processo onde o que estava contido na mensagem do discurso *web* era a base a partir da qual se faziam inferências e tiravam conclusões sobre o conteúdo. Este tipo de análise permitiu representar

formalmente o conteúdo das trocas discursivas, com referência aos significados, contextos e intenções contidas nas mensagens.

Para analisar os conteúdos do discurso *web*, Herring (2013) propôs uma alteração do conceito que designou de *web content analysis* de forma a abarcar todo o conteúdo das RSO, como temas, recursos e *links*. Segundo a autora, como todos podem servir para comunicar, logo são passíveis de ser analisados. Todavia, como referido em Freire et al. (2017), recursos não linguísticos que não sejam baseados em caracteres, tais como ":-)", ";(", etc., requerem um pré-processamento e conhecimento específico de outras áreas de conhecimento.

No discurso *web*, os utilizadores equipararam-se aos oradores e aos ouvintes que comunicam entre si através das RSO. Os atos de fala das trocas discursivas foram vistos como sendo as mensagens trocadas entre utilizadores. Tomando como exemplo o *Facebook*, as mensagens são designadas por *post* e *comment*. Os *posts* iniciam uma conversação e os *comments* são as respostas. No caso do *Twitter*, as mensagens são designadas como *tweet*s. O conteúdo das mensagens, recursos linguísticos e não linguísticos, são o que cada utilizador expressa, não oralmente, mas sim na forma escrita utilizando as tecnologias da *web* social.

Olhar para o discurso *web* através das teorias clássicas do discurso permitiu analisar o comportamento das RSO através da perspetiva da linguagem, onde os seus significados são fundamentados em observações da utilização da língua e numa perspetiva linguística. Esta perspetiva refletiu a aplicação de conhecimentos e metodologias oriundas do estudo da linguagem falada e escrita, tais como a análise da conversação, a análise crítica do discurso e, dentro da análise de conteúdo, a análise semântica de textos. Assim, a comunicação e interação existente entre utilizadores foi recolhida a partir das trocas discursivas produzidas no discurso *web*. Essas trocas discursivas estabelecem-se através das relações de trocas comunicativas onde, segundo Gee (2001), os intervenientes criam, quando falam ou escrevem o que têm a dizer, para se ajustarem à situação ou contexto onde comunicam.

Esta abordagem permitiu também a formalização dos conceitos importantes para a análise do conteúdo semântico do discurso *web*, enquanto recursos discursivos específicos. Numa análise semântica é importante definir cada conceito para que possa ser codificado de forma coerente, ou seja, para definir o significado de uma palavra dentro de um discurso específico. Desta forma evitou-se a subjetividade e ambiguidade que podia existir em torno de um conceito, visto que este podia assumir significados diferentes, dependendo do contexto onde era utilizado.

A análise do discurso *web* remete também para o PLN porque são necessários conhecimentos e técnicas da ciência da computação para processar, codificar e explorar os conteúdos aí produzidos. A área do PLN, segundo Jurafsky e Martin (2009), designa-se também por discurso do computador e processamento da linguagem, tecnologia da linguagem humana ou linguística computacional. Por si só, o PLN é uma área multidisciplinar que envolve a inteligência artificial, os SI e as TI, e a linguística (Antunes et al., 2014). O PLN utiliza métodos computacionais, suportados em modelos e linguagens formais, para o tratamento da linguagem humana, desenvolvendo sistemas com a capacidade de reconhecer e de reproduzir a informação da linguagem natural.

Embora as linguagens formais sejam artificiais, quer estas, quer a linguagem natural têm estruturas semelhantes, o que permite a aplicação da lógica na semântica do discurso *web*. De acordo com Jurafsky e Martin (2009), o objetivo do PLN é conseguir que os computadores executem tarefas que envolvam a linguagem humana, permitindo a comunicação homem-máquina, melhorar a comunicação entre as pessoas e processar texto ou voz. Para Russell e Klassen (2019), o PLN permite analisar um documento, constituído por um conjunto ordenado de símbolos (que segue uma sintaxe própria e uma gramática razoavelmente bem definida) e infere a semântica associada a esses símbolos.

Na análise do discurso *web*, a análise de conteúdo e as técnicas de PLN são só uma das peças do puzzle que utilizam o texto como *input* de dados, categorizam e padronizam o texto, atribuindo números às categorias ou a partes do texto e, em seguida, calculam estatísticas. A análise do discurso *web*, no seu todo, tem o seu foco não só no significado das mensagens, no valor atribuído a cada pedaço de texto, mas também nos seus intervenientes.

2.2 Redes sociais *online*

O conceito de rede social tem diversos significados na literatura e é percebido através de diferentes pontos de vista, consoante a área de investigação. De acordo com Freeman (2004), a abordagem da rede social fundamenta-se na noção intuitiva de que a padronização dos vínculos sociais entre os utilizadores tem consequências importantes para estes. Para Hanneman e Riddle (2005), as redes sociais são, antes de mais, relações de interação entre indivíduos, sejam elas em causa própria, em defesa de outros ou em nome de uma organização. As interações de indivíduos nas suas relações quotidianas, sejam elas familiares, comunitárias, de amizade, trabalho ou estudo, entre outros, caracterizam as redes sociais informais que surgem de forma espontânea. Todavia, as redes sociais também podem

surgir de forma intencional, sendo criadas e divulgadas por indivíduos, organizações ou grupos com poder de liderança, que articulam pessoas em torno de interesses, projetos ou objetivos comuns. Dentro destas últimas destacam-se as *FanPages* criadas pelas empresas para manterem uma presença institucional *online*.

Nas ciências sociais, segundo Izquierdo e Hanneman (2006), a rede social define-se geralmente como um conjunto (grande ou pequeno) de atores sociais que podem ter relacionamentos uns com os outros e um ou mais tipos de vínculos entre eles. De igual modo, Jamali e Abolhassani (2006) definem o conceito de RSO como sendo um conjunto de pessoas, organizações ou outras entidades sociais, interligadas por um conjunto de relações sociais, tais como amizade, trabalho ou troca de informação. Segundo Boyd e Ellison (2007), os utilizadores pretendem, em primeiro lugar, comunicar com as pessoas que fazem parte da sua rede social, porque estas oferecem fóruns de discussão, grupos ou outros recursos de comunicação que partilham os mesmos interesses.

Kaplan e Haenlein (2010) definem *media* sociais como as aplicações da *Web*, construídas a partir do fundamento ideológico e tecnológico da *Web* 2.0, que permitem a criação e troca de conteúdo produzidos pelos utilizadores. Por seu lado, para Dziczkowski et al. (2010), na sociedade contemporânea, o termo rede social utiliza-se para descrever um grupo de atores sociais que interagem através de um qualquer meio de comunicação *online*. A estas definições, Marmo (2011) acrescenta que as redes sociais são estruturadas. Segundo alguns autores (Antunes e Costa, 2011; Boyd e Ellison, 2007), as plataformas de redes sociais são ambientes onde os indivíduos criam perfis auto descritivos e, de seguida, os divulgam através de *links* para outras pessoas que conhecem, criando deste modo uma rede de ligações pessoais.

Turban et al. (2018) definem *media* sociais como uma plataforma de conteúdos *online* de texto, imagem, áudio e vídeo criados pelo utilizador, acessível através das ferramentas da *Web* 2.0. Para os autores, os *media* sociais tem como finalidade a interação e conversas sociais, tais como partilha de opiniões, experiências, conhecimento e colaboração *online*. Savic et al. (2019) utilizam a designação redes sociais para descreverem interações e relações entre entidades sociais, que podem ser indivíduos, grupos, instituições ou até nações inteiras.

Como estabelecido em Freire et al. (2021), as designações *media* sociais *online* e rede social confundem-se com frequência, sendo importante estabelecer as suas diferenças. Os *media* sociais *online* são plataformas de comunicação *online*, que têm por base a interação

social. Em contraste, o conceito rede social é muito mais antigo do que as tecnologias digitais da *Web*. As redes sociais têm origem nos estudos realizados nas áreas da sociologia e sociometria, que investigam as interações existentes entre as pessoas.

No entanto, com o desenvolvimento das tecnologias digitais, as redes sociais foram associadas aos *media* sociais *online* porque o foco da sua análise é sobretudo a *web* social, a sua estrutura de rede, a interação entre utilizadores e/ou a conteúdo semântico que produzem. Neste trabalho de investigação, para interligar os *media* sociais *online* e a rede social, utilizou-se o conceito RSO, que abarca quer as tecnologias digitais, quer os relacionamentos entre as pessoas. Nesse sentido, uma RSO é um ambiente *online* onde os utilizadores partilham informação, como por exemplo o *Facebook*, o *Twitter*, o *WhatsApp*, entre outros.

As RSO têm facultado o registo de múltiplos aspetos da natureza humana relacionados com o modo como comunicamos, quer em termos pessoais, quer em termos organizacionais. Como refere a literatura, estas redes permitem uma maior interação entre pessoas e/ou organizações (utilizadores da *web* social), levando a que estes utilizadores se expressem e sejam seguidos por audiências restritas ou globais. Gyarmati e Trinh (2009) corroboram a ideia de que as RSO mudaram a forma como os seres humanos se relacionam e entram em contacto uns com os outros. Como referem Antunes e Costa (2012b), as tecnologias de RSO estão também a mudar a forma como as pessoas e as organizações interagem e partilham a informação.

De facto, este novo paradigma acaba por funcionar como um meio de comunicação de eleição, a que as empresas não devem ficar alheias, visto que ele pode alterar consideravelmente aspetos da rotina organizacional. A comunicação tem-se modificado devido à velocidade com que a informação é transmitida a quem se encontra ligado a uma qualquer interface, independentemente do tempo e do espaço.

As pessoas adotam e integram as tecnologias nas suas rotinas diárias, cada vez com maior frequência (Bechmann e Lomborg, 2015; Isson, 2018; Kanagavel, 2019). Isto significa que, hoje em dia, independentemente do equipamento utilizado e através das RSO é possível uma interação direta e ativa entre utilizadores, incrementando assim as trocas discursivas próprias do discurso *web*.

Este comportamento, de partilha de informação e opiniões entre utilizadores, desencadeia a acumulação de uma quantidade colossal de informação (Dey et al., 2019; Kim et al., 2017; Missaoui et al., 2017; Moreira et al., 2019; Savic et al., 2019; Xhafa et al., 2015),

e, implicitamente, há uma necessidade acrescida de métodos de análise dessa informação, para determinar opiniões e pontos de vista sobre diversas questões. Essa informação, depois de processada, pode desempenhar um papel importante na tomada de decisão nas organizações.

2.3 Aspetos relevantes do discurso *web*

O discurso *web* é de uma riqueza tal que deve envolver níveis e métodos de análise tanto das ciências sociais (ligadas à interação entre indivíduos), como da ciência da computação (ligadas à formalização e interpretação). Isto significa que a análise integral do discurso *web* é inevitavelmente uma tarefa multidisciplinar e interdisciplinar e que a sua complexidade obriga a que se façam escolhas específicas entre os vários métodos disponíveis para o efeito, dependendo do objeto de estudo e funções a analisar. Independentemente de toda a complexidade que envolve a análise do discurso *web*, existem pontos comuns nas diversas formas para o analisar. Todas são unânimes em considerar que um discurso é um conjunto de significados, através dos quais um grupo de pessoas comunica sobre um determinado tema. Nesse sentido, como referido em Freire et al. (2015b) para a análise semântica das mensagens produzidos no discurso *web*, é importante perceber aspetos tais como: (i) a estrutura da RSO; (ii) a comunicação; (iii) o contexto; (iv) a linguagem; (v) a linguística, semântica e pragmática e (vi) os conteúdos.

No atinente à estrutura da RSO (i), é importante perceber as inter-relações estruturais que compõem a rede quer dos utilizadores, quer das mensagens que trocam, no contexto da análise do discurso *web*. Neste âmbito, as técnicas da SNA permitem perceber a estrutura do discurso, identificando as interações que se estabelecem entre os utilizadores e evidenciando os vínculos sociais. As métricas da SNA desempenham, assim, um papel importante na identificação de utilizadores-chave e das interações entre eles, pois quando apresentam valores elevados indicam prestígio e influência, independência e capacidade de controlar e difundir informação. Estas métricas permitem perceber as inter-relações que surgem numa RSO em linguagem quantitativa e visual, para assim explicar o comportamento dos utilizadores e quantificar o seu poder enquanto oradores capazes de persuadir o seu público-alvo.

A SNA trouxe, então, um sentido analítico para o conceito de rede, permitindo estudar a sua estrutura social a partir de uma perspetiva relacional, colocando no centro da investigação o elemento principal da sociologia: a interação social (Duijn e Vermunt, 2006;

Wasserman e Faust, 1994). Os aspetos estruturais das redes permitem perceber a estrutura global do discurso *web*, identificar as relações estabelecidas entre as várias entidades de análise e evidenciar os vínculos sociais (Shum et al., 2011) que possam existir entre utilizadores e que facilitam a troca de informação (Jamali e Abolhassani, 2006).

O segundo aspeto, a comunicação (ii), encontra-se relacionado com a própria estrutura organizativa das RSO, ou seja, as topologias de rede. Aqui evidenciam-se dois pontos de análise importantes. Primeiro, a capacidade de aglomeração e/ou a dispersão dos utilizadores, que pode indicar o poder da RSO (nesse sentido, o entendimento da topologia da RSO permite não só às organizações definirem as suas formas de participação *online*, mas também identificar produtores de informação e evidenciar o papel que tal produtor exerce no discurso *web*) e, segundo, a análise da frequência de mensagens trocadas e análise do conteúdo dessas mensagens (que está relacionado com a dinâmica de construção e formação da RSO). As topologias da rede, sobre a partilha de conteúdos e trocas discursivas, traduzem o processo de comunicação que normalmente está estruturado por critérios de afinidade, interesses e/ou conhecimento dos seus utilizadores.

As trocas discursivas e a sequência com que ocorrem têm uma relação direta com o tipo de comunicação, podendo ser vistas como sendo síncronas ou assíncronas. A comunicação síncrona tem resposta imediata, pelo que as interações decorrem num período temporal próximo, de modo análogo à conversação presencial, ou seja, em tempo real. Nesse tipo de comunicação, elementos como as interações, as trocas discursivas e os períodos temporais, (características denominadas discursivas como em (Van Dijk, 1985)) são mais evidentes e facilmente identificáveis, visto que geralmente aparecem seguidos. Todavia, na maioria dos casos, apesar das interações serem síncronas, um único utilizador pode comunicar em simultâneo com vários outros. Esta situação pode dificultar a identificação da sequência das interações e estruturação da comunicação do discurso *web*. Por isso, a análise da sequência de um ato comunicativo pode ser auxiliada pela análise do ato seguinte.

Já a comunicação assíncrona é aquela em que a expectativa de resposta não é imediata e as interações acontecem dentro de um período temporal mais alargado, pelo que, a sua análise pode ser mais complexa. Para compreender essas estruturas e sequências de interação, é preciso identificar os respetivos pares conversacionais, ou seja, como é que as mensagens estão relacionadas umas com as outras e a que utilizador pertence cada uma. Essa identificação, no âmbito deste trabalho, permitiu não só perceber quem falava com quem, como também quando o faz. Para efeitos de formalização, a compreensão da estrutura dos

pares conversacionais auxiliou diretamente a compreensão da sequência das interações. Se por um lado a identificação do tipo de comunicação associada às trocas discursivas permitiu compreender a sua estrutura e a das interações, por outro, a sua identificação também pôde ser construída com base no texto contido na mensagem de cada utilizador.

Independentemente da comunicação entre utilizadores se estabelecer tanto de forma síncrona como assíncrona, esta pode apresentar características únicas tais como *smiles*, siglas, abreviaturas e vocabulário especiais (Georgalou, 2010). Como refere Herring (2013), a maior parte do discurso *web* é constituído por texto, áudio, vídeo e/ou imagens estáticas. Se por um lado esta conjugação de recursos linguísticos e não-linguísticos para comunicar proporciona uma maior fluência nas trocas discursivas, por outro lado, este tipo de comunicação desencadeou uma nova forma de linguagem híbrida entre a linguagem escrita e a falada, utilizadas a partir de contextos criados pelos utilizadores. Deste modo, e segundo Fairclough (2003), considera-se a utilização da linguagem como um ato comunicativo, que tem associado um propósito, onde estão subjacentes regras específicas do contexto específico onde está a ser produzido o discurso.

O terceiro aspeto diz respeito ao contexto (iii) onde é produzido o discurso *web*, visto que a sua formalização e análise não passam apenas pelo entendimento da linguagem, mas também pelos aspetos contextuais onde ocorrem as trocas discursivas. Segundo Locke (2004), é o contexto social que torna significativos os vocabulários, a sintaxe e a estrutura do discurso. Os aspetos sociais de um contexto reportam à própria cultura que, de acordo com Boyd e Ellison (2007), são diversos e emergem em torno das RSO, a que Beer (2009) e Lytras et al. (2009) chamam de participativas. Como referem Castrucci et al. (2011), o contexto define o modo pelo qual as interações se concretizam dentro de cada cultura, sendo a soma da identidade e do comportamento individual dos diversos participantes, que ajuda a definir a sua identidade social.

Assim, em cada contexto, as trocas discursivas têm associada uma carga de subjetividade e ambiguidade, justificadas por Perelman e Olbrechts-Tyteca (2005) pelas experiências de partilha e construção das diferentes realidades dos intervenientes. No seio das RSO, esta subjetividade e ambiguidade desenvolvem-se entre dois ou mais utilizadores através de uma comunicação linguística escrita, característica do discurso *web*, constituída por linguagem natural, elementos verbais e não verbais, mas também do contexto. Para além disso, as tecnologias digitais têm vindo a aumentar ainda mais a complexidade das trocas

discursivas, visto que as tornaram mais diversificadas, rápidas e, consequentemente, de difícil formalização e codificação para posterior observação.

Para formalizar e codificar a complexidade das trocas discursivas foi necessário considerar várias questões, pois as RSO são dinâmicas e, portanto, em constante alteração. Uma delas, foi considerar o que Van Dijk (1977) designa de contexto atual, isto é, um período de tempo e um lugar onde os acontecimentos entre falante e ouvinte se desenrolam, ou seja, o aqui e agora, de forma lógica e cognitiva. Com base nessa definição, delimita-se formalmente um contexto como sendo um desenrolar de acontecimentos que têm um estado inicial, vários estados intermédios e um estado final.

Outra questão que se teve em conta foi o facto de que, dependendo do contexto, a linguagem natural permite que os intervenientes do discurso utilizem diferentes designações para se referirem à mesma entidade (Jurafsky e Martin, 2009). Por exemplo, quando falamos do nosso carro é possível designá-lo de formas diferentes. Dependendo do contexto onde ocorre o discurso é possível utilizar as expressões "este carro", "o carro", "o Volkswagen", "o meu Polo" ou "o meu carro", entre outras possibilidades. A escolha entre as várias alternativas não é livre e independente do contexto. Não se pode simplesmente dizer "aquele" ou o "Volkswagen", pois o ouvinte pode não ter conhecimento prévio do carro que temos ou o mesmo pode não ter sido mencionado antes. Em sentido oposto, a linguagem natural permite a utilização da mesma designação para referir entidades diferentes. O conceito, em inglês, *seat* utilizado por uma companhia aérea tem um significado diferente do que se for utilizado por um fabricante automóvel. Para a primeira, o significado é literalmente um assento. Já para o segundo, significa a marca de um carro. De acordo com Jurafsky e Martin (2009), isto define o contexto situacional do discurso.

Assim, em termos computacionais, é importante determinar, definir e compreender o significado dos conceitos dentro do contexto em que surgem, isto é, dentro do contexto específico. Como a interpretação do discurso *web* pode ser um processo impreciso, os significados devem ser entendidos como propriedades da interação entre as palavras e os contextos. Só com uma análise da estrutura do discurso *web* (gramatical, semântica, etc.), não é possível reconstruir de forma completa o que foi expressado. Este aspeto é sobretudo importante na codificação de dados semânticos para posterior utilização em processamentos automáticos. Russell e Klassen (2019) reforçam esta ideia quando declaram que, mesmo tendo informação perfeita acerca da estrutura de uma frase, ainda é preciso um contexto adicional, exterior à mesma, para a interpretar corretamente. Nesse sentido, para a

reconstrução semântica das mensagens produzidas no discurso *web*, considerou-se que a informação podia ser codificada para em seguida a adaptar a um contexto específico.

Segundo Herring (2013) e Russell e Klassen (2019), a análise do contexto da *web* social deve considerar não só o lugar onde a interação acontece e os utilizadores, mas também o objetivo do discurso, um vez que em cada contexto a interação social dos utilizadores pode ser motivada e/ou alterada consoante o significado ou interpretação dos símbolos (palavras, *smiles*, etc.) utilizados no discurso. Tal significa que, se por um lado o contexto influencia os utilizadores a produzirem mais conteúdos no discurso *web*, por outro lado, esses mesmos conteúdos podem descrever os utilizadores através de dados dinâmicos, homogéneos e semânticos inerentes a um determinado contexto.

Devido ao seu carácter dinâmico, o contexto é fundamental para a análise do discurso *web*, pois este altera-se a cada instante à medida que os utilizadores interagem uns com os outros. Todavia, também é abstrato (subjetivo e ambíguo), visto que está sujeito aos fatores socioculturais, à capacidade que os utilizadores têm para criar novos conceitos e às respetivas emoções. O contexto do discurso envolve, para cada interveniente, as suas convicções, crenças, pressupostos partilhados (ou não) com os outros, as sua intenções e expectativas, as relações que se estabelecem e as representações mentais que cada um tem de tudo isso (Van Dijk, 1985). Nesse sentido, a análise de um contexto específico, para apoio à decisão organizacional, deve delimitar-se a toda a informação definida como relevante para responder a uma questão específica e não a toda a informação existente na *web* social.

Relativamente à linguagem (iv), Locke (2004) refere que as linguagens sociais além de dependerem do contexto onde são utilizadas, também utilizam uma gramática própria. No contexto das RSO as linguagens são muitas, pois são o resultado da utilização das tecnologias, sendo suscetíveis a adaptações e a mudanças de acordo com a necessidade de cada utilizador. Como referem Pozzi et al. (2017), grande parte da linguagem escrita está agora nos ecrãs dos computadores, *tablets* e *smartphones*, evoluindo, em parte, devido à interação dos utilizadores com a tecnologia. De acordo com Jokinen e Wilcock (2017), se por um lado os novos paradigmas tecnológicos alteram a utilização da linguagem, por outro, ela também se adapta às alterações tecnológicas.

Assim, para expressarem os novos conceitos que emergem nas RSO, bem como para comunicarem de uma forma mais ágil, os utilizadores utilizam e misturam recursos linguísticos e não-linguísticos (Bodomo, 2010). De acordo com Gee (2001), a linguagem tem uma propriedade "mágica" pois, quando se fala ou escreve, cria-se o que se pretende

dizer para se ajustar à situação ou contexto onde se comunica. Para Perelman e Olbrechts-Tyteca (2005), qualquer símbolo pode servir de meio de comunicação, desde que integrado numa linguagem compreendida pelos ouvintes. Wooffitt (2005) considera que os símbolos podem assumir diversos e/ou diferentes significados de acordo com o seu contexto. Este facto reforça a ideia de que, inserindo-os no discurso *web*, o utilizador tem liberdade para atribuir e/ou alterar o significado das palavras de acordo com as suas intenções, experiências e representações.

A linguagem do discurso *web* é deste modo específica em muitos aspetos, o que lhe confere algumas características que devem ser consideradas numa análise semântica. O que mais sobressai é a formação de novas palavras, visto que um grande número de siglas e abreviaturas emergiram recentemente. A linguagem utilizada no discurso *web* caracteriza-se também pela tendência de reduzir a quantidade necessária de caracteres que se digitam para expressar uma ideia e, consequentemente, para acelerar o processo de comunicação (Antunes et al., 2018; Freire et al., 2015b). Neste âmbito, normalmente, na linguagem escrita, a pontuação é utilizada para separar unidades gramaticais, tais como frases e/ou conceitos (Jurafsky e Martin, 2009)**.** Todavia, e de acordo com Bodomo (2010), é interessante notar que novas formas de utilizar a pontuação têm sido desenvolvidas pelos utilizadores da *web* social. A utilização de sinais de pontuação tais como "!!!!", "????" e/ou "!!??" têm o objetivo de enfatizar uma ideia que se pretende passar na mensagem.

Outro aspeto que caracteriza o discurso *web* é a emoção, visto que os intervenientes posicionam o seu discurso em relação a si mesmos, evidenciando as suas próprias posições (a favor ou contra) e as suas emoções. Captar essas emoções, numa matéria inerte como são os dados semânticos (texto), é de extrema dificuldade porque as trocas discursivas ocorrem através da interação, que é complexa, e da convergência de vários recursos e estratégias linguísticas. Segundo Locke (2004), na comunicação presencial as emoções podem facilmente ser comunicadas através de pistas não-verbais. Como referem Morville e Rosenfeld (2006), quando se comunica com outra pessoa no discurso presencial, conta-se com o retorno (*feedback*) constante do interlocutor (por exemplo com a linguagem corporal) para ajudar a aperfeiçoar a forma como a mensagem é transmitida. No entanto, quando se comunica através da linguagem escrita, nomeadamente no discurso *web*, é difícil exprimir e transmitir emoções com símbolos ou palavras. Uma das formas de colocar emoção no texto é utilizar uma combinação de caracteres do teclado que representam uma expressão facial

ou utilizar símbolos predefinidos (*smiles*) existentes nas aplicações que se utilizam para comunicar.

Se por um lado os recursos linguísticos (texto) são predominantes no discurso *web*, por outro, os recursos não linguísticos (código visual ou semiótico[19]), tais como imagens e/ou vídeos em substituição de palavras ou até mesmo de frases (Herring, 2013), são cada vez mais utilizados. Nesse sentido, os textos produzidos no âmbito do discurso *web* podem ser considerados multi-semióticos porque, na sua construção, combinam diferentes formas de representação da linguagem. Isto deve-se ao facto de os utilizadores terem vindo a desenvolver formas de articular os vários tipos de recursos linguísticos, transportando-os para o discurso *web* para tornar a comunicação *online* similar à comunicação presencial. Esta nova forma de comunicar no discurso *web* privilegia uma linguagem informal, em oposição a uma linguagem mais formal e estruturada.

A linguagem informal utiliza-se no dia-a-dia no discurso verbal (presencial) e não exige muita atenção por parte da gramática, de forma a que haja mais fluidez na comunicação. Também se caracteriza por utilizar gíria e conceitos, que na linguagem formal não estão padronizados ou têm outro significado, sendo de difícil formalização e interpretação computacional. Em sentido oposto, a linguagem formal carrega consigo a rigidez das normas gramaticais, utilizando-se sobretudo em textos e domínios específicos (por exemplo no meio académico), sendo por isso de mais fácil formalização e interpretação computacional.

A linguagem utilizada e produzida nas RSO tenta simular o discurso presencial que é um processo dinâmico e, quando passa para a forma escrita, pode ser transformada em dados, ou seja, algo que pode ser capturado, processado e analisado. As suas características reforçam a ideia de que a linguagem utilizada para comunicar na *web* social tende a ser mais maleável do que a escrita formal (Pozzi et al., 2017). Isto leva a que, por norma, os dados semânticos do discurso *web* sejam na sua maioria desordenados, incompletos e, por vezes, com erros, caracterizando-se por serem não estruturados e informais (Antunes et al., 2018). Todavia, como refere a literatura (Osei-Bryson e Rayward-Smith, 2009; Pozzi et al., 2017; Russell e Klassen, 2019) quando utilizados corretamente e bem explorados, são um recurso

[19] A semiótica, enquanto ciência, utiliza símbolos para representar factos quotidianos, ou seja, estuda o mundo das representações enquanto linguagem (Morris, 1938).

valioso para a gestão das organizações, uma vez que podem acrescentar conhecimento à organização.

O quinto aspeto (v) encontra-se associado à linguística, semântica e pragmática. Do ponto de vista da semântica e da pragmática, a linguagem caracteriza-se por ter um significado para alguém, num contexto específico, com um determinado objetivo onde são utilizados símbolos combinados entre si através de regras. De acordo com Morville e Rosenfeld (2006) e Dale (2010), o código da linguagem natural, para além dos símbolos, inclui a gramática (sintaxe), a semântica e a relação pragmática entre eles.

No atinente aos símbolos, estes são unidades fundamentais em qualquer idioma e, na sua forma escrita, constituem o alfabeto (Jurafsky e Martin, 2009). Assim, o discurso *web* (texto) articula-se através do vocabulário (ou seja, do conjunto de símbolos que transportam o significado) e pela sintaxe (pelo conjunto de regras que combinam esses símbolos de forma a terem significado). Deste modo, a sintaxe diz respeito à disposição e à combinação dos símbolos dentro das frases (proposições) e das frases dentro do discurso.

Relativamente à semântica, de acordo com Morris (1938) e Fairclough (2003), esta preocupa-se com a forma como os símbolos se combinam entre si, de forma a criarem uma comunicação com significado. Nas RSO as ideias contidas no discurso *web* expressam-se através de frases que são escritas, utilizando símbolos que projetam uma determinada situação ou realidade. Do ponto de vista da semântica, cada conceito existente numa frase tem associado um objeto que representa uma realidade. Essa realidade pode ser inferida e formulada através da forma lógica que existe na frase. De acordo com Gartner (2016), a semântica de um discurso é uma propriedade que caracteriza a combinação do significado das palavras. Nesse sentido Pozzi et al. (2017) reforçam que a semântica da linguagem utilizada nas RSO é fundamental para analisar com precisão as expressões dos utilizadores.

A codificação semântica dos conteúdos das trocas discursivas envolve necessariamente uma componente subjetiva de interpretação que, para Herring (2004), na maioria dos casos só pode ser realizada por programadores humanos e não inferida por agentes de *software*. Para Wachsmuth (2015), é necessária a intervenção humana, porque a criação de algoritmos depende da informação que se vai encontrar. O conhecimento (linguagem) do contexto é capturado manualmente e transferido do especialista (humano) para a máquina onde, como referido em Freire et al. (2017), o envolvimento humano é essencial para analisar sistematicamente os resultados obtidos e verificar se são necessários ajustes na codificação.

Segundo Mika (2007), na área da computação, capturar e analisar dados do contexto social permite que as máquinas, através de agentes, possam aplicar regras de inferência (raciocinar). No contexto da *web* semântica e de acordo com Berners-Lee et al. (2001), os computadores precisam de ter acesso a informação estruturada (dados e metadados[20]) e a conjuntos de regras de inferência, que auxiliem o processo de dedução automática, para que se consiga fazer o raciocínio de forma automática, ou seja, representar o conhecimento. Estas regras de inferência especificam-se através de ontologias, que permitem representar de forma explícita a semântica associada aos dados. Assim, através das ontologias é possível construir e representar redes de conhecimento humano que complementam o processamento das máquinas e, portanto, melhorar significativamente os serviços da *web*.

No entanto, as ontologias apresentam alguns problemas, independentemente das potenciais vantagens. Apesar das ontologias terem como objetivo estruturar, de forma organizada, a informação de uma área de conhecimento, refletindo um entendimento semântico de situações reais e automatizar a comunicação entre pessoas e máquinas, não conseguem abarcar todos os conceitos e termos de uma linguagem em particular. Para além disso, verifica-se que o significado de uma mensagem é muito mais do que o conteúdo semântico associado às suas palavras (Bodomo, 2010; Herring, 2013; Russell e Klassen, 2019), pois **e**m cada contexto existe uma cultura com características próprias e uma forma diferente de se exprimir, ou seja, uma linguagem própria. Quer isto dizer que, se um contexto tem uma semântica própria, logo são necessárias ontologias associadas a essa semântica em particular.

A complexidade linguística dos conteúdos das RSO ainda não permite utilizar de forma simples o conhecimento existente no domínio da *web* semântica. A própria literatura (Jurafsky e Martin, 2009; Russell e Klassen, 2019) salienta que a codificação e interpretação de alguns dados têm algumas dificuldades. Estas são provocadas pela ambiguidade semântica dos conceitos que podem ter associadas diferentes representações de significado em diferentes contextos, com base nas circunstâncias em que a comunicação ocorre. Esta ambiguidade pode trazer restrições à análise das relações existentes entre partes de texto ou conceitos do discurso *web*. Por esse motivo, vários autores (Gartner, 2016; Jurafsky e Martin, 2009; Russell e Klassen, 2019) reforçam a ideia de que o papel do contexto também é importante nos casos em que é necessário decidir se duas palavras têm significados

[20] Metadados são dados referentes a outros dados que facilitam o entendimento da informação e obtenção de conhecimento (Gartner, 2016; Riley, 2017).

semelhantes. Nestes casos, as palavras com significados análogos utilizam-se, na maioria das vezes, em contextos semelhantes, quer em termos de estrutura sintática da frase, quer com definições semelhantes no dicionário.

Os significados trocados nas interações comunicativas da *web* social não só são uma questão relacionada com a semântica, mas também com a pragmática. A pragmática é o ramo da linguística que estuda a linguagem no contexto de comunicação onde é utilizada, onde as palavras podem assumir significados distintos. Para Morris (1938) e Wooffitt (2005), a pragmática é definida como sendo um ramo da linguística que estabelece a relação dos símbolos com os seus intérpretes. Para Van Dijk (1977), a pragmática tem por base a linguagem, a teoria dos atos de fala (trocas discursivas), a análise da conversação e as diferenças culturais existentes na interação. Segundo Fairclough (2003), a unidade de análise da pragmática não é a frase em si mesma, nem a sua enunciação por parte do orador, mas sim, os jogos de linguagem e as várias formas como são utilizados. De forma similar, para Jurafsky e Martin (2009), a pragmática está para além da construção da frase, que é estudada pela sintaxe, ou do seu significado, estudado pela semântica, visto que estuda essencialmente os objetivos da comunicação.

De acordo com Bharti et al. (2017) e Kreuz e Caucci (2007), as características pragmáticas do discurso *web* incluem não só o texto, mas também todos os outros recursos não linguísticos, tais como *emoticons*[21], *smiles*, caracteres de pontuação, *replies*, entre outros. Assim, a pragmática tem o seu foco nos significados linguísticos determinados não só pela semântica, mas também por aqueles que se deduzem a partir de um contexto discursivo de uma situação ou acontecimento. Isto significa que a semântica está dependente da pragmática porque os interlocutores também dependem dos pressupostos ou expectativas sobre o que está a ser dito, de modo a inferirem o significado da mensagem.

Nesse sentido, a pragmática utilizou-se na análise de RSO para apoiar a transformação das trocas discursivas, posicionando-as dentro de uma situação e definindo em que condições eram produzidas. A pragmática foi, por isso, importante para a formalização e análise do discurso *web*, no sentido de dar resposta a problemas práticos, desdobrando os fenómenos linguísticos de acordo com diferentes pontos de vista e entendimentos.

[21] *Emoticons*: estratégias linguísticas para comunicar estados emocionais, tais como vogais repetidas, sinais de pontuação repetidos, sequência de caracteres que representam expressões faciais, entre outros (Kappas e Kramer, 2011).

Contudo, a semântica e a relação pragmática são muito difíceis de formalizar, sendo por norma consideradas como pertencentes ao domínio da inteligência artificial (Morville e Rosenfeld, 2006). Fazer a distinção entre as relações semânticas e as relações pragmáticas facilitou a compreensão e tratamento da linguagem, especificando como cada segmento de texto devia ser interpretado em relação a outro. Para além disso, permitiu perceber as relações de coerência do discurso *web* e complementar a semântica com a pragmática, como refere a literatura (Morris, 1938; Van Dijk, 1977). Para Saint-Dizier (2012), identificar no discurso estas relações requer diferentes tipos de conhecimento e de raciocínio, visto que algumas situações são extremamente difíceis de resolver, enquanto que outras podem ser processadas por inferência lexical ou de raciocínio efetuado sobre conhecimento ontológico.

O sexto e último aspeto (vi) diz respeito aos conteúdos produzidos nas RSO, pois a sua análise permite compreender os processos nos quais os utilizadores se envolvem ao trocarem mensagens. De acordo com Lai e To (2015) e Laudon e Laudon (2011), o conteúdo criado pelo utilizador é uma das características que definem a *web* atual. De igual modo, Antunes e Costa (2012b) consideram que as contribuições dos utilizadores são o elemento-chave do conteúdo da *web* social. As RSO alteraram o paradigma de como a informação está a ser produzida, transferida e consumida, visto que através da *internet* e das tecnologias da *web* esta está mais acessível (Symeonidis et al., 2014).

A análise do discurso *web* pode realizar-se a partir de um nível mais global de cima para baixo, ou seja, começar a sua interpretação, identificando que tipos de recursos estão contidos no discurso, ou de baixo para cima, onde o ponto de partida é o detalhe linguístico. De acordo com Herring (2010), a análise de conteúdo é uma técnica estruturada que abarca a codificação e classificação dos símbolos encontrados na comunicação, as suas características estruturais e os seus aspetos semânticos. O conteúdo pode, por isso, ser classificado por quantidade e tamanho das mensagens, distribuição de certos elementos de texto, imagem, etc. Os conteúdos das RSO são atualmente uma das áreas de análise mais importante, visto que englobam não só o que as pessoas dizem, mas também o que partilham. De acordo com Golbeck (2015) e Russell e Klassen (2019), o conteúdo das mensagens, ou seja, o texto, as fotos e os vídeos, bem como a quantificação quer de *likes*, quer de visualizações, são o que de mais valioso se pode encontrar numa investigação em RSO.

Tudo o que os utilizadores colocam nas RSO pode ser transformado em dados estruturados e, por isso, quantificáveis: quer o tipo de mensagens e características, quer o que contêm. Como refere Felt (2016), as mensagens (*posts*) das RSO estão cheios de potencial

para mineração e análise de dados. Estes dados traduzem o que as pessoas andam a fazer, as suas opiniões, com quem interagem e porque o fazem. Segundo Osei-Bryson e Rayward-Smith (2009), se forem bem utilizados e explorados, podem ser um bem valioso para a gestão das organizações. De acordo com Lai e To (2015), este tipo de informação pode ajudar, por exemplo, profissionais de *marketing* a monitorizar as perceções das pessoas e ajudar as organizações no planeamento estratégico.

De acordo com Moreira et al. (2019), os textos são a forma mais comum de trocar informação na nossa sociedade. No entanto, o conteúdo produzido na *web* social, enquanto meio de comunicação, não está confinado à utilização de recursos linguísticos, ou seja, texto. As novas tecnologias oferecem meios cada vez mais sofisticados para a interação social, onde se comunica não apenas utilizando texto, mas também através de recursos não-linguísticos. Este novo tipo de discurso, que combina formatos e sistemas de codificação, envolve novos padrões de utilização e colaboração do utilizador na sua produção e interpretação (Herring, 2013).

Como refere Golbeck (2015), as pessoas podem colocar um *post* sobre qualquer coisa nas RSO. Dentro do discurso *web*, a criação de sentido da comunicação, formado por esta linguagem consegue-se através da articulação e inserção de alguns (ou todos) estes recursos. Logo, no conteúdo das mensagens, e de acordo com Bodomo (2010), Herring (2013), Jurafsky e Martin (2009) e Laudon e Laudon (2011), podemos encontrar uma mistura de texto, imagens estáticas (fotos, captura de outros conteúdos), vídeos, áudio (música, voz e quais quer outro tipo de som), *tags*[22], *links* e símbolos (s*miles*, pontuação, etc.).

Recolher e analisar os diferentes recursos linguísticos do conteúdo da *web* social requer diferentes abordagens. Para uma interpretação mais ampla, cada recurso linguístico ou não-linguístico deve ser analisado como uma unidade independente (Herring, 2010). A interpretação em unidades independentes permite uma descrição objetiva, sistemática e quantitativa do conteúdo da comunicação da *web* social. Embora o conteúdo de uma imagem possa ser analisado tendo em conta os seus temas e características, a interpretação do conteúdo visual beneficia da aplicação dos métodos da semiótica. Por exemplo, os dados de imagem exigem *software* específico para entender o conteúdo da imagem e traduzi-lo para

[22] As *tags* permitem categorizar e descrever o conteúdo com palavras-chave que podem ser utilizadas como termos de pesquisa.

conteúdo textual (ver Bouet et al. (2009), Ignatov et al. (2017) e Pennington (2017), para aprofundamento deste assunto).

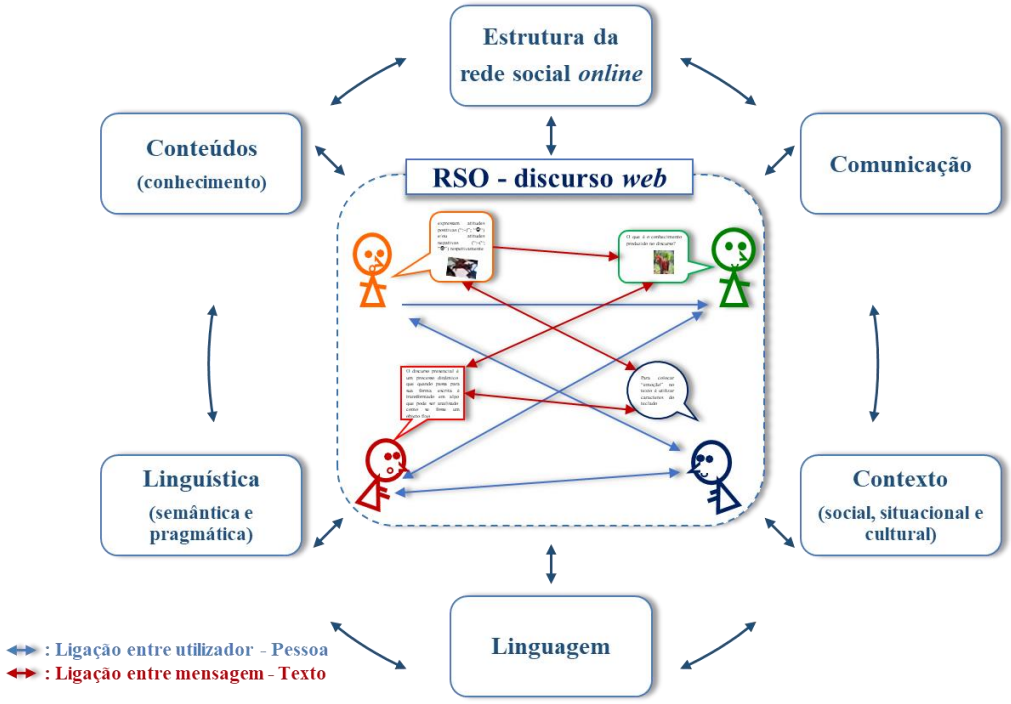
O conteúdo, enquanto variável do discurso *web*, pode ser visto na perspetiva dos métodos para extração e análise de dados, bem como na perspetiva do conhecimento, enquanto produto final das interações sociais. Na perspetiva dos métodos de extração e análise, como refere Aufaure et al. (2006), a mineração *web* pode extrair padrões de dados através da mineração de conteúdos, mineração da estrutura e mineração de utilização da *web*. A mineração de conteúdos é, assim, uma forma de exploração de texto aplicada às mensagens das RSO. Relativamente à mineração da estrutura, a mesma utiliza-se para analisar os dados da estrutura da RSO onde o principal foco são as ligações que unem as mensagens. A mineração de utilização aplica-se aos dados de utilização da *web* pois, segundo Markov e Larose (2007), estes refletem o comportamento dos seres humanos e a sua interação na *web*.

No âmbito da análise do discurso *web*, a mineração foi útil por dois motivos: para além de servir para perceber as interações entre utilizadores, serviu também para adicionar novos conceitos nas estruturas semânticas utilizadas para padronizar o texto das mensagens.

É de salientar que a mineração, por si só, se divide em *data-mining* e *text-mining*. A primeira utilizou-se com o objetivo de procura padrões escondidos e com significado dentro de bases de dados, enquanto a segunda para procurar informação em fontes não estruturadas, mais concretamente em dados registados de forma textual e numa linguagem comum.

A Figura 8 representa os aspetos considerados importantes para análise do texto produzido no âmbito do discurso *web*. Este ocorre dentro de uma estrutura de RSO onde os intervenientes comunicam entre si (interagem), dentro de um contexto que utiliza uma linguagem própria que deve ser analisada tendo em conta aspetos linguísticos para se perceber o conteúdo produzido. Assim, no centro foi colocada a RSO constituída pela rede de utilizadores e rede de texto, ligada a cada um dos aspetos. Esta ligação pretende representar, por um lado, que todos os seis aspetos estão interligados e dependem uns dos outros e, por outro, que qualquer aspeto do discurso *web* deve ser entendido tendo em conta a interação intrínseca à rede de utilizadores.

Figura 8 - Aspetos importantes do discurso *web*



2.4 Entidades de análise do discurso *web*

Este trabalho de investigação apoiou-se na teoria dos grafos e nas técnicas da SNA para representar a configuração da RSO, pois rede e grafo são conceitos sinónimos. Todavia, a representação formal da complexidade de uma rede é constituída unicamente por dois elementos, os nós e os seus vínculos (Robinson et al., 2015; Savic et al., 2019; Wasserman e Faust, 1994; Zheng e Skillicorn, 2017), onde os diversos modelos e diagramas mostram configurações diversas, sempre através de pontos e linhas. Para o efeito, as dimensões que os dados das RSO representam são as ligações entre os três níveis em análise. Todavia, para construir um modelo de análise de dados de RSO foi necessário identificar alguns elementos, tais como o tipo de entidades a analisar e as ligações entre elas (vínculos).

Pela literatura constatou-se que o conceito RSO era formulado a partir de definições que remetiam para a comunicação e/ou troca de informação e vínculos de interação, não hierárquicos, entre utilizadores. Verificou-se também que na literatura existiam vários conceitos para designar as várias entidades em análise - ator social, agente, nó, ponto ou vértice, das RSO e outros tantos para designar os vínculos entre elas - relações, *links*, laços, linhas ou ligação.

De acordo com as definições das ciências sociais, o discurso, enquanto meio de comunicação, tem três entidades essenciais: o orador; os atos de fala e os recursos linguísticos e não linguísticos. De modo a estabelecer um paralelismo entre as teorias clássicas e o discurso *web*, no modelo desenvolvido, os utilizadores das RSO assumiram a figura de oradores, as mensagens definiram-se como sendo um conjunto de atos de fala e os conteúdos como sendo os recursos linguísticos e não linguísticos utilizados.

Assim, o discurso *web* ficou constituído por três entidades distintas a analisar e interligar: o utilizador (*user*); a mensagem (*post*) e o conceito (*concept*). O objetivo foi utilizar estas três entidades para transpor formalmente o discurso *web* das RSO. As várias abordagens da análise do discurso ajudaram a explicitar formalmente a estrutura discursiva, quer ao nível da interação entre os utilizadores, quer ao nível dos conteúdos que produziam. Para além disso, auxiliaram a representação do discurso *web* em unidades de análise, de modo a facilitar a sua codificação e criação de conjuntos de dados.

Este trabalho de investigação definiu a entidade *utilizador* como sendo qualquer pessoa individual ou coletiva que produz conteúdos nas RSO. Nesse sentido, a entidade *utilizador* pode analisar-se através da SNA, não só para estudar os participantes e os seus antecedentes, mas também para obter uma visão geral e ampla das características dos principais participantes de um grupo. A entidade *utilizador* desempenha, assim, um papel importante na identificação de utilizadores-chave e das interações entre eles. A sua análise permite explicar o seu comportamento e quantificar o seu poder enquanto disseminadores de informação.

Na entidade de análise *mensagem* englobaram-se todas as mensagens e consequentes respostas e/ou comentários produzidos pelos utilizadores no âmbito das suas interações. As mensagens foram classificadas, quantificadas e analisadas como uma única unidade de acordo como o seu tipo e características, tais como os seus recursos linguísticos e não linguísticos.

A entidade *mensagem*, quando associada às sequências da comunicação, ou seja, as trocas discursivas, torna-se um fator importante para a análise do discurso *web*. Isto significa que, para se compreenderem as trocas discursivas, foi necessário perceber como se relacionavam entre si. O vínculo de relacionamento é identificável pelos elementos da comunicação, data e hora das interações e pela designação que cada RSO dá à sua primeira mensagem e consequentes comentários. Por exemplo, o *Facebook* designa a primeira mensagem que deu origem a cada troca discursiva de *post* e as respostas/comentários a este

último designa de *comments.* Já o *Twitter* designa as mensagens de *tweets*, não fazendo distinção entre a primeira e as seguintes. Cada RSO tem a sua denominação própria para designar uma resposta e/ou comentário a uma mensagem. A sua sequência permite compreender que troca discursiva vem primeiro e com que interação está relacionada. Deste modo analisaram-se os aspetos estruturais do discurso através dos elementos da comunicação e os semânticos através da análise do conteúdo das trocas discursivas. Desta forma foi possível compreender quem falava com quem e como se encontrava organizado o discurso *web*.

A entidade de análise *conceito* englobou todos os recursos linguísticos textuais contidos nas mensagens produzidas pelos utilizadores no âmbito das suas interações. A análise das mensagens, no que diz respeito aos *conceitos*, foi importante para classificar o detalhe linguístico. Isto é, transformar em dados cada pedaço de texto contido nas trocas discursivas, para a realização de análises quantitativas e/ou qualitativas. De acordo com Pozzi et al. (2017), a quantificação "do que as pessoas pensam" é uma etapa indispensável para que os dados textuais sejam verdadeiramente úteis para os processos de decisão.

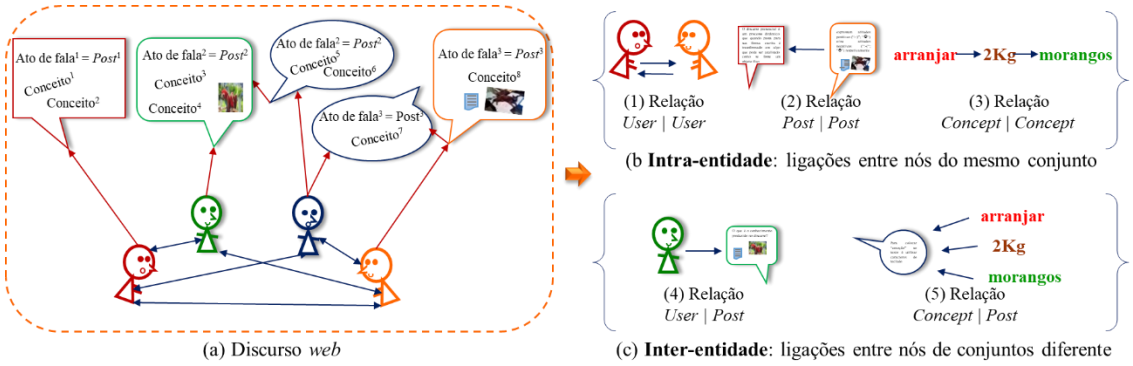
Sabendo que dentro dos conteúdos das trocas discursivas é possível encontrar conceitos importantes e úteis para o apoio à decisão, esses conceitos podem ser estruturados e identificados com as métricas da SNA. Deste modo, determinou-se que a estrutura das trocas discursivas era formada pelos recursos linguísticos da linguagem, na forma de texto escrito. O método de criação das redes semânticas assentou no entendimento de que cada mensagem podia ser convertida em unidades de texto, no sentido de as resumir de duas formas: rede semântica de resumo de todas as unidades de texto da mensagem e rede semântica de palavras-chave. O pressuposto subjacente foi que quer uma quer outra descreviam o conteúdo textual das trocas discursivas. De acordo com alguns autores, (Aggarwal, 2011; Pang e Lee, 2008), extrair e analisar conceitos individuais, palavras-chave, pode fornecer um resumo de uma opinião expressa no discurso *web* e permitir uma extração mais refinada da mesma.

A esta estrutura semântica podem juntar-se os outros dois elementos do discurso *web*: utilizadores e mensagens. Isto permitiu que quer os resumos, quer as palavras-chave das trocas discursivas fossem diretamente ligadas a um utilizador. Assim, o foco não fica só na análise das interações entre utilizadores, mas também na análise da estrutura da comunicação e do que era dito por eles a cada momento.

No que se refere às ligações entre utilizadores, a estrutura da rede do discurso *web* permitiu considerar os fatores que ligavam o emissor ao seu recetor e, tanto a origem, como o destino da mensagem ajudaram a definir como surgiam as interações, importando saber não só quem estava e para quem estava a "discursar", como também em que contexto era produzida e recebida a informação.

Depois da identificação das entidades, a segunda unidade a identificar e a caracterizar foi o tipo de ligação existente entre elas. Para o efeito, foi necessário identificar os vínculos entre a mesma entidade (*intra*) ou entre diferentes entidades (*inter*). Como evidenciado em (a) na Figura 9, o utilizador verde é "amigo" do vermelho e do laranja pelo que, neste caso, estamos perante interação entre utilizadores que pertencem todos à entidade definida como utilizador. O utilizador azul escreveu uma mensagem que foi comentada pelo utilizador laranja. Neste caso, estamos perante uma troca discursiva em que, para além de uma interação entre utilizadores (*user* | *user*), temos uma ligação estabelecida entre duas mensagens (*post* | *post*) e os utilizadores que as escreveram (*user* | *post*). A mensagem do utilizador laranja tem um conteúdo textual. Neste caso temos uma ligação estabelecida entre dois ou mais conceitos (*concept* | *concept*) e uma ligação entre a entidade *concept* e a entidade *post* (*concept* | *post*). Na Figura 9, em (b) e (c), caracterizam-se todas as possíveis ligações inter e intra entidade do discurso *web*.

Figura 9 - Discurso *web* e caracterização das relações inter e intra-entidade



Neste âmbito, as técnicas da SNA permitem uma análise das ligações e da respetiva comunicação entre utilizadores, do seu fluxo de informação, das relações estabelecidas entre eles, bem como da sua influência noutros indivíduos, grupos ou entidades. De acordo com Jamali e Abolhassani (2006), a SNA proporciona uma análise quer visual, quer matemática das interações humanas e trocas de comunicação. Foi a partir da estrutura da rede, composta pelas três entidades de análise, que foi possível estabelecer vínculos entre o processo de

comunicação e o respetivo discurso *web* como um todo, bem como as suas aplicações, tal como no apoio à decisão.

2.5 Resumo

Neste capítulo, sistematizaram-se os conceitos de redes sociais e RSO que se encontram interligados entre si. Se, por um lado, a *web* social tem como motor para o seu funcionamento a colaboração, a partilha e o fluxo de informação, características das redes sociais. Por outro, a sua estruturação formal, tem por base os nós, entidades centrais de análise e as ligações (interações entre utilizadores). Quando juntamos o modo como funcionam as RSO e as bases que as constituem, percebe-se que existem questões centrais que efetivam o processo de comunicação inerente à *web* social. Deste modo, para analisar o discurso *web* é importante perceber como se formam e como se desenvolvem as ligações entre utilizadores e qual o seu papel enquanto produtores de informação.

Neste capítulo examinaram-se também as definições tradicionais de discurso, ligadas às ciências sociais, uma vez que são necessárias para formalizar os conteúdos no contexto das ciências da computação. Assim, foram abordadas a forma restrita, a forma ampla e a retórica da análise do discurso. A forma restrita envolve a semântica, a forma ampla está associada às interações entre utilizadores e a retórica está ligada a artifícios de linguagem e ao contexto organizacional.

As relações entre as mensagens e a interpretação do sentido do que é trocado remetem para a análise do discurso (por si só uma área multidisciplinar). Levando em consideração que o conteúdo do discurso *web* é complexo, pois tem associada a argumentação que se suporta de jogos de linguagem característicos dos ambientes digitais, apontaram-se várias perspetivas de análise do discurso *web*, nomeadamente a análise da conversação visto que esta estuda os elementos do discurso (utilizador, mensagem, conceitos) e que o texto contido nas mensagens, do discurso *web*, pode ser visto como uma sequência de atos de fala. Cumulativamente, examinou-se quer a análise critica do discurso, para estudar o contexto em que o discurso *web* é produzido, quer a análise de conteúdo, para estudar a semântica contida discurso *web*.

Os conhecimentos teóricos associados ao estudo das RSO, enquanto forma de comunicação onde ocorrem trocas discursivas, foram, assim, centrais neste capítulo. Para o efeito, foi importante identificar e perceber os aspetos que indicam como observar o discurso

web, que emerge das trocas discursivas produzidas na *web* social e em particular nas RSO. Também se identificaram os aspetos associados ao discurso *web* tais como a estrutura da rede de utilizadores, a comunicação, contexto, linguagem, linguística e conteúdo. Estes aspetos foram importantes, pois a *web* social reflete estruturas sociais que são construídas e modificadas pelos seus intervenientes através de ferramentas de comunicação onde se incluem as RSO. Essas RSO são constituídas por grupos de indivíduos cujas trocas discursivas geram vínculos e conhecimento, onde se estabelecem comunicações síncronas e assíncronas que influenciam de forma diferente a estrutura da rede social. Tendo em conta que dentro dos diversos discursos, as palavras e frases estão sempre carregadas de sentidos e intenções, que podem ser definidas de forma diferente, foi necessário ter em consideração o contexto no qual estas apareciam inseridas. Também houve que ter em conta que na análise do discurso *web*, porque a compreensão da linguagem se estende para além do seu formalismo gramatical e se incorporam as intenções do orador (que utiliza determinadas palavras para atingir os seus objetivos), foi necessário perceber o que era a perspetiva pragmática.

Por último, foram analisadas as três entidades de análise do discurso *web*, visto que a sua representação formal recorre ao conhecimento da teoria dos grafos e das técnicas SNA.

3 Social network analysis (SNA)

O conceito de rede tem vindo a ser formulado em diferentes domínios a partir de designações que remetem para a interação entre indivíduos e vínculos não hierárquicos, tendo todos em comum relações de comunicação e/ou troca de informação. De acordo com a literatura (Freeman, 2004; Wasserman e Faust, 1994), os primeiros estudos de representação das interações entre indivíduos, que levaram ao desenvolvimento da SNA, tiveram origem na década de 30. Essas representações, usualmente têm a forma de redes *one-mode* ou grafos, em que todos os nós pertencem a um mesmo conjunto, isto é, os indivíduos que interagem entre si. Contudo, a SNA tem vindo a ser aplicada em muitas áreas e as redes deixaram de ser compostas só por indivíduos, para passarem a integrar outros elementos ou entidades de análise.

Como referido no capítulo anterior, o discurso *web* é constituído por três entidades distintas a interligar e analisar: o *utilizador*, a *mensagem* e o *conceito*. Estudar as estruturas das RSO remete para a teoria dos grafos e para a SNA. Por definição, a teoria dos grafos é um ramo da matemática e da ciência da computação que lida com estruturas matemáticas, permitindo modelar pares de relações entre entidades (Burguillo, 2018; Diestel, 2017; Grohe, 2017).

Em termos gerais, pode referir-se que um utilizador está ligado à entidade mensagem, se a interação do utilizador produziu uma troca discursiva. Pode-se ainda dizer que a entidade mensagem está ligada à entidade conceito, se à mesma estiverem associados recursos textuais. Partindo dessa premissa, o discurso *web* pode formalizar-se através de redes *two-mode* que são compostas por subconjuntos de utilizadores que participam em atividades sociais. Como as atividades sociais, normalmente, são constituídas por várias entidades, em vez de exclusivamente pares de utilizadores, uma rede *two-mode* é formada com a informação das sub-redes referentes a cada entidade.

A análise de uma rede *two-mode* permite uma compreensão mais rica das interações humanas e dos seus interesses do que a obtida como uma rede *one-mode*. Através da recolha de informação sobre as ligações existente entre as três entidades de análise, isto é, *utilizador*, *mensagem* e *conceito*, é possível ter uma boa imagem de toda a rede de utilizadores, bem como das mensagens que trocam. As redes *two-mode* permitem agregar as três entidades de análise de RSO e a SNA ajuda na obtenção de informação útil sobre a estrutura da rede e

dos padrões de relacionamento, tais como quem fala com quem e que conteúdo foi transmitido.

De acordo com a literatura (D'Andrea et al., 2010; Erétéo et al., 2011; Leech et al., 2018; Moreira et al., 2019; Wasserman e Faust, 1994), as métricas da metodologia SNA são geralmente apropriadas para vários níveis de análise da rede e a interpretação dos seus resultados pode subdividir-se de acordo com três níveis diferentes: (i) interpretação de toda a rede; e (ii) interpretação de conjuntos e componentes (*socio-centric*); (iii) interpretação das posições individuais dos atores (*ego-centric*) na rede.

Para Laumann et al. (1992), numa investigação em redes sociais a primeira questão a especificar é a definição dos seus limites, pois são esses limites que vão determinar a amostra-alvo. Portanto, na interpretação do discurso *web* das RSO, só interessa dividir a análise estrutural em *ego-centric* e *socio-centric*, para a delimitar ao contexto onde este se encontra inserido. Analisar toda a rede não faz sentido, pois esta é constituída por múltiplos contextos. Assim, dentro da análise de RSO destacam-se as dimensões *ego-centric* e *socio-centric*, representadas na Figura 10, ambas de grande importância para estudar as interações entre utilizadores e para construir redes semânticas.

Os investigadores de SNA têm vindo a desenvolver um conjunto de conceitos e métricas específicas para a análise de RSO (Mika, 2007). Como ilustra a Figura 10, as métricas da metodologia SNA podem assumir um âmbito global ou individual, isto é, se estão centradas na estrutura global da rede ou se estão centradas nos nós da mesma.

A primeira dimensão, *socio-centric*, é uma análise de rede de âmbito global utilizada para investigar conjuntos de vínculos entre nós delimitados em termos coletivos e sociais (D'Andrea et al., 2010; Marsden, 2002; Scott, 2000). A segunda dimensão, *ego-centric*, é uma análise de rede de âmbito individual e centra-se num nó específico e não na rede como um todo (Duijn e Vermunt, 2006; Hanneman e Riddle, 2005; Leech et al., 2018). As redes *ego-centric* dão uma visão da rede, a partir da perspetiva de um único nó e das suas ligações.

Figura 10 - Dimensões de análise da SNA

De acordo com a literatura (Abraham et al., 2010; Carrington et al., 2005; Hanneman e Riddle, 2005; Marmo, 2011; Ortiz-Arroyo, 2010; Scott, 2000; Wasserman e Faust, 1994), existem propriedades e métodos associados à análise quer *socio-centric*, quer *ego-centric*. As propriedades e métodos associados apenas para a análise *ego-centric* incluem a forma como os nós se destacam dentro de um grupo e são quantificadas através de métricas como a centralidade e prestígio, o seu nível de expansividade e os seus parâmetros de popularidade. Existem ainda métricas para funções individuais, que identificam o quanto o nó se encontra isolado, os seus contactos e pontes. Por outro lado, podem utilizar-se métricas aplicáveis a pares de nós (*socio-centric*) e às ligações existentes entre eles. Para o efeito, a teoria dos grafos tem métricas que evidenciam a distância e a proximidade entre nós, noções estruturais e outras de equivalência que postulam modelos estatísticos para os diversos fins e tendências probabilísticas, relacionadas com a reciprocidade.

Associados às dimensões que permitem estudar as redes em função do volume de nós, na SNA existem vários conceitos chave (Wasserman e Faust, 1994). Estes conceitos,

representados na Figura 11, definidos e descritos na literatura (Alhajj e Rokne, 2018; Savic et al., 2019; Scott, 2000; Wasserman e Faust, 1994), são fundamentais na análise de RSO:

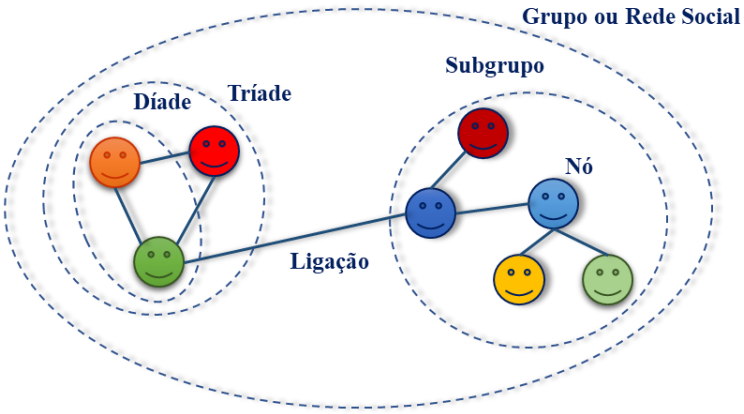
- Ator (*Actor*) - De uma forma abrangente, os atores são entidades sociais que estão relacionadas entre si. Estas entidades sociais são de vários tipos, mais concretamente, um indivíduo, uma organização, um grupo ou subgrupo, uma mensagem, um texto, entre outros. Este conceito permite diferentes níveis de agregação, podendo utilizar-se em diferentes problemas de investigação. Na literatura, este conceito é designado de formas diferentes, tais como: ator social, agente, nó, ponto, vértice. Assim, para designar cada uma das três entidades de análise das RSO, neste trabalho de investigação utilizou-se a designação "nó";
- Laço relacional (*Relational tie*) - Um nó, qualquer que seja a sua entidade, encontra-se ligado a outro através de um laço relacional, responsável por estabelecer um vínculo entre um par de nós. Estes laços existem entre nós que estão ligados por uma ou mais relação. Também para este conceito, na literatura, existem várias designações, tais como: vínculo, relação, *link*, linha, laço, ligação. De forma a evitar alguma ambiguidade no significado do conceito, utilizou-se o termo "ligação", para designar os vínculos entre os nós;
- Subgrupo (*Subgroup*) - É um subconjunto de nós e todas as ligações existentes entre eles. Este subconjunto pode ser uma díade que consiste num par de nós e as ligações existente entre eles. A análise das díades tem o seu foco nas propriedades da relação dos pares, nomeadamente se as ligações são recíprocas ou não, ou se ligações múltiplas e específicas tendem a ocorrer em conjunto. O subconjunto pode também ser uma tríade formada por três nós e as ligações estabelecidas entre eles, ou ainda ser constituído por um conjunto finito de mais de três nós;
- Grupo (*Group*) - É o conjunto finito de nós e de todas as ligações que, por razões conceituais, teóricas ou empíricas, estão a ser medidas e analisadas. Na análise de RSO existe um grupo ou conjunto que corresponde a cada entidade de análise, ou seja, existe um grupo para os utilizadores, um para as mensagens e outro para os conceitos;
- Relação (*Relational*) - Conjunto específico de ligações entre nós de um qualquer grupo. O tipo de relação é definido de acordo com as ligações entre os nós. Nas RSO, as ligações são explícitas e declaradas quando presentes entre indivíduos,

mensagens ou conceitos. As ligações entre indivíduos designam-se por interações entre utilizadores, entre mensagens de trocas discursivas e entre conceitos por resumos semânticos;

- Rede social (*Social Network*) - Consiste num conjunto finito ou grupo de nós e a relação ou relações existentes entre eles. As redes sociais, quer sejam *online* ou não, normalmente estruturam-se de acordo com as ligações existentes entre indivíduos. No entanto, podem estruturar-se, analisar-se e estudar-se ligações que correspondem a grupos, organizações, nações ou até mesmo palavras.

Os parâmetros principais de análise das RSO são as métricas da SNA. No entanto, e como referem Hanneman e Riddle (2005), da panóplia de técnicas matemáticas disponíveis na SNA, não existem formas nem indicadores certos ou errados na abordagem às RSO. Por seu lado, Brandes e Erlebach (2005) salientam que, dependendo da rede em questão, podem ser utilizadas diversas métricas. No âmbito da aplicabilidade da SNA para PLN, Banati et al. (2017) referem que são muitas as possíveis combinações de métricas para identificar padrões interessantes nos dados. Em cada caso deve ser selecionado um conjunto de técnicas, de acordo com os objetivos do estudo, as características estruturais e a dinâmica da rede a analisar.

Figura 11 - Representação dos conceitos chave da SNA



3.1 Estrutura da rede e métricas globais

As métricas que ajudam a descrever a estrutura global da rede, descritas na literatura (Burguillo, 2018; Carley, 2015; Hanneman e Riddle, 2005; Izquierdo e Hanneman, 2006; Moreira et al., 2019; Savic et al., 2019; Scott, 2000; Wasserman e Faust, 1994), são:

- *Diâmetro* - Por definição é a maior distância geodésica[23] da rede e indica o número de passos que são necessários para ir de um nó a outro. Ou seja, o comprimento do caminho mais curto que pode ligar qualquer par de nós na rede. O diâmetro é por vezes utilizado como uma medida de conectividade de rede;
- *Tamanho* - O tamanho da rede significa o número de nós ligados ou, em alternativa, o número de nós na rede. O tamanho da RSO pode ser crucial para a análise e estruturação das relações sociais, devido aos recursos por vezes limitados de armazenamento e processamento das ligações entre os nós;
- *Densidade* - A densidade é definida como sendo o número de ligações existentes na rede, expressa através da percentagem do número de todos os vínculos possíveis. Ou seja, é o número de ligações reais da rede, dividido pelo número de todos os que possam estar presentes e que de facto estão. A densidade da rede dá uma perspetiva da velocidade a que flui a informação, a dimensão dos conteúdos produzidos e/ou restrições sociais dos nós, sendo também por vezes utilizada como uma medida de conectividade da rede. A análise da densidade é importante na identificação de comunidades (*clusters*), pois estas são constituídas por estruturas mais densas, compostas pelas ligações existentes entre nós. A importância de uma rede está na capacidade que esta tem de estabelecer ligações dentro e fora dela. Por outras palavras, a importância da rede está diretamente associada à quantidade de ligações que os nós estabelecem entre si e não ao número de nós que a compõem;
- *Distância* - A distância entre dois nós é o número mínimo de ligações necessárias para passar de um para outro, sendo também conhecida como a distância geodésica, representando a sua proximidade. A distância entre os nós de uma rede, representa o quanto eles estão perto um do outro. Se, por um lado, em distâncias grandes a informação demora mais tempo a chegar e a ser difundida para toda uma rede, por outro, distâncias curtas permitem a disseminação mais rápida da informação. Para além disso, os nós mais próximos podem exercer uma maior influência, uns nos outros, do que sobre outros que estão mais distantes.

[23] Distância geodésica é o número mínimo de laços necessários para passar de um nó para outro (Izquierdo e Hanneman, 2006).

Tão ou mais importante do que as métricas centradas na estrutura global da rede, são as métricas centradas nos nós da rede, ou seja, nas características de cada elemento. Todavia, o cálculo da posição estratégica ou centralidade de um elemento, no que se refere à sua atividade, controlo, importância, influência, poder, prestígio e proximidade, depende da definição de centralidade utilizada.

3.2 Posição estratégica das entidades da rede

Se, por um lado, a estrutura da rede afeta a sua tendência evolutiva, por outro, o próprio processo evolutivo, que implica a adição continuada de novos elementos, pode modificar a rede. Segundo Hanneman e Riddle (2005), a questão de como a posição estrutural caracteriza um elemento (e lhe confere importância e poder), continua a ser um tópico de investigação ativo e de debate considerável. Os investigadores têm vindo a desenvolver um número significativo de métricas de centralidade, cujo objetivo é medir a variação da importância dos nós de acordo com critérios predefinidos (Abraham et al., 2010; Newman, 2005; Salamanos et al., 2017). Essa problemática, da identificação da posição estratégica e importância de elementos da rede, também designados por utilizadores-chave, continua a atrair muitos investigadores, como mostram Mojtaba et al. (2019).

De acordo com a literatura (Alhajj e Rokne, 2018; Izquierdo e Hanneman, 2006; Scott, 2000; Vasconcelos e Barão, 2017; Wasserman e Faust, 1994), os elementos mais importantes estão, normalmente, localizados em posições estratégicas da rede. Um elemento é importante se as suas ligações o tornam particularmente visível para os outros elementos da rede. Ou seja, a sua importância caracteriza-se pela sua centralidade e pelo seu prestígio, o que lhe confere poder dentro da rede. Elementos importantes são alvo de muitas escolhas por parte de seguidores e são aqueles que estão muito envolvidos no relacionamento com outros. Assim, do ponto de vista das definições de centralidade, um elemento é central quando está envolvido em muitas ligações. Do ponto de vista do prestígio, *status* ou popularidade, um elemento tem prestígio quando é o nó de entrada das ligações. Esse prestígio aumenta à medida que o elemento é alvo de mais ligações, mas não necessariamente quando ele próprio inicia essas ligações. Por outras palavras, é preciso olhar para as ligações direcionadas a um elemento da rede para analisar o seu prestígio.

No âmbito da análise de RSO, calcular a centralidade de um elemento, isto é, de um nó, tem como objetivo identificar a sua posição estratégica, relativamente às trocas discursivas que faz e às ligações que estabelece dentro da rede. Como acima referido, o

conceito de centralidade traz associada a ideia de poder, conferido pela posição estratégica e pelo prestígio. As diferentes definições e métricas podem interpretar de forma diferente qual o nó de origem do poder, bem como podem representar, de forma diferente, as estruturas sociais.

A literatura referente às métricas de centralidade é extensa, como se pode ver por exemplo em Bonacich (1987, 2007), Everett e Borgatti (2005), Freeman (1979), Friedkin (1991), Lozares et al. (2015) e Newman (2005). O trabalho de Freeman (1979) abordou o conceito de centralidade, analisando um conjunto de métricas já publicadas, reduzindo-as a três definições clássicas: *closeness centrality* (centralidade de proximidade), *betweenness centrality* (centralidade de intermediação) e *degree centrality* (centralidade de grau). A este conjunto de métricas de centralidade consideradas importantes, vários autores (Golbeck, 2015; Ortiz-Arroyo, 2010) acrescentaram mais uma: a *eigenvector centrality*, também apenas designada por *eigenvector*. Estas quatro métricas de centralidade são, assim, as mais utilizadas para avaliar a posição estratégica de um nó específico na rede. A literatura produzida, quer nas ciências sociais, quer nas ciências da computação, bem como a aplicação sistemática das métricas de centralidade da SNA em trabalhos de investigação nestes últimos 15 anos, por si só validam a sua importância. Marsden (2002), justifica que a utilização das métricas de centralidade se deve à perceção e à importância atribuídas aos conceitos de poder social e influência estrutural.

A métrica *closeness centrality*, por definição, mede o quanto um nó está próximo dos restantes, enfatizando a sua distância ou proximidade relativamente aos outros dentro da rede (Burguillo, 2018; Freeman, 1979; Hanneman e Riddle, 2005; Moreira et al., 2019; Wasserman e Faust, 1994). Para Izquierdo e Hanneman (2006) a *closeness centrality* representa independência associada à possibilidade de comunicação com muitos nós da rede, através de um número mínimo de intermediários. No entanto, segundo os autores, existem várias definições que dependem do significado que cada um dá à expressão "estar perto". Para Burguillo (2018), a métrica *closeness centrality* reflete a habilidade ou capacidade que um indivíduo tem em disseminar a informação para outros elementos da rede, de forma rápida.

A noção de proximidade baseia-se no conceito de distância mínima ou geodésica $d(n_i, n_j)$, ou seja, o número mínimo de vínculos necessários para ir do nó n_i para o nó n_j. De acordo com Freeman (1979) e Wasserman e Faust (1994) a *closeness centrality* (C_c) do

nó n_i é dada pelo inverso da soma das distâncias de n_i a todos os outros nós, onde N é o número total de nós, sendo formalizada por:

$$C_c(n_i) = \left[\sum_{j=1}^{N} d(n_i, n_j)\right]^{-1}, \text{com i, j } \in \text{ N} \wedge \text{i, j, N} \in \mathbb{N} \quad (1)$$

A métrica *betweenness centrality*, por definição, tem subjacente a ideia de que um nó é central quando está entre muitos nós e isso torna-o poderoso (Freeman, 1979; Wasserman e Faust, 1994). Existe assim a assunção implícita de que os caminhos mais curtos são avaliados pela *betweenness centrality* (C_B) e que um caminho mais curto tem maior probabilidade de ser escolhido relativamente a outro. Quanto mais central é um indivíduo, maior a capacidade para controlar o fluxo de informação, recursos e comunicações, que circulam entre os nós (Izquierdo e Hanneman, 2006). A *betweenness centrality* está relacionada com a capacidade que um elemento tem para ser um intermediário entre quaisquer outros dois elementos da rede.

Assim, segundo Wasserman e Faust (1994), assumindo que g_{jk} é o número de ligações geodésicas que ligam dois nós n_j e n_k e $g_{jk}(n_i)$, o número de ligações geodésicas entre os nós n_j e n_k que passam por n_i, a *betweenness centrality* do nó n_i pode definir-se como:

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}, \text{com i, j } \in \text{ N} \wedge \text{i, j, N} \in \mathbb{N} \quad (2)$$

Também para Newman (2005) e Burguillo (2018) a definição convencional de *betweenness centrality* mede a influência que um nó tem sobre a disseminação da informação na rede, calculando apenas os caminhos mais curtos. Como refere Golbeck (2015), esta métrica é amplamente utilizada para identificar pessoas influentes na transmissão de informação de uma parte da rede para outra. Implicitamente, assume-se que a informação se dissemina apenas ao longo desses caminhos mais curtos. Partindo desta premissa, e com base nos percursos aleatórios, Newman (2005) introduziu uma nova métrica designada de *random walks betweenness centrality* (centralidade de intermediação de percursos aleatórios). Esta inclui contribuições de praticamente todos os caminhos entre nós, não apenas os mais curtos, apesar de dar mais peso a estes. A métrica assenta em caminhos aleatórios, contando quantas vezes um nó é atravessado por um percurso aleatório entre dois outros nós.

Na metodologia SNA, uma métrica fundamental e amplamente utilizada é a *degree centrality* de um nó (Antunes e Costa, 2012a; Everett e Borgatti, 2005; Mika, 2007; Vasconcelos e Barão, 2017). De acordo com a literatura (Alhajj e Rokne, 2018; Brandes e Erlebach, 2005; Freeman, 1979; Savic et al., 2019; Scott, 2000; Wasserman e Faust, 1994), a métrica *degree centrality* mede a importância de um nó, quantificando o número de ligações entre dois nós específicos. Esta métrica, para além de quantificar as ligações diretas de um nó, descreve quantos elementos compõem a sua vizinhança e a forma como a informação flui entre eles.

A definição mais simples da métrica *degree centrality* firma-se na ideia de que os nós importantes são os mais ativos, no sentido em que possuem um maior número de ligações a outros nós. Por definição, na literatura, a *degree centrality* de um nó é o número de nós adjacentes, onde dois nós são considerados adjacentes quando existe um vínculo entre eles. De acordo com Wasserman e Faust (1994), a *degree centrality* do nó n_i, denotado por C_D, é o número de vínculos do nó n_i. Segundo Freeman (1979), R_{ji} são elementos da matriz adjacente, onde $R_{ji} = 1$, se e só se os nós n_i e n_j tiverem um vínculo entre si, caso contrário é 0 e pode definir-se como:

$$C_D(n_i) = \sum_{j=1}^{N} R_{ji} \text{ , com i, j } \in \text{ N} \wedge \text{i, j, N} \in \mathbb{N} \quad (3)$$

Para além da definição anterior, a *degree centrality* também indica se os nós têm ou não um sentido definido. O sentido da ligação entre os nós remete para os conceitos de grafo *undirected* (não direcionado) e *directed* (direcionado). Numa rede *undirected*, o grau é o número de ligações de um nó e representa apenas a sua presença ou ausência e não diz nada sobre a força das mesmas (Erétéo et al., 2011; Golbeck, 2015; Grohe, 2017; Izquierdo e Hanneman, 2006; Ozyer et al., 2019; Savic et al., 2019; Zheng e Skillicorn, 2017). Em sentido oposto, as redes *directed*, identificam a força de cada vínculo, pois aos mesmos são atribuídos valores. Assim, nas redes direcionadas, é necessário fazer a distinção entre ligações de entrada e de saída, traduzidas pelas métricas *in-degree* e *out-degree*, respetivamente.

O conceito de *out-degree* é uma medida que usualmente indica quão influente o nó pode ser dentro da rede, ou seja, evidencia a atividade que um elemento tem ao responder aos outros, por exemplo, quando o utilizador envia informações para outros. Podemos também ver os elementos da rede como recetores de informação, isto é, o quanto outros enviam informações ou mantêm ligações com um determinado nó. O facto de um nó receber

muita informação de outros pode indicar prestígio ou poder, ou então sofrer daquilo a que se designa como fadiga de informação. O *in-degree* de um ponto é mostrado pela soma das colunas na matriz do grafo, enquanto que o seu *out-degree* é mostrado pela soma das linhas.

Sintetizando, nas redes direcionadas o *degree centrality* é o número total de extremidades ligadas a um nó e existem duas métricas possíveis: o *out-degree* que é o número de ligações com origem num nó e o *in-degree* que é o número de ligações que apontam para um nó (Burguillo, 2018; Erétéo et al., 2011; Hansen et al., 2011; Ozyer et al., 2013; Savic et al., 2019). Nas redes direcionadas, considerar a direção dos nós pode alterar a interpretação da métrica *closeness centrality* acima referida. Se a origem e o destino das ligações representam, respetivamente, a capacidade de alcançar um elemento ou ser alcançado por outros, então os elementos da rede têm prestígio quando estão próximos de outros e, em simultâneo, são nós de entrada (Wasserman e Faust, 1994). Ou seja, quando têm valores elevados de *closeness centrality* e de *in-degree*. Quando têm um valor elevado de *closeness centrality* e um valor elevado de *out-degree*, é considerado um elemento de suporte (Wasserman e Faust, 1994). A ideia é que um nó é central quando interage rapidamente com os outros e não apenas com os que se encontram diretamente ligados a ele.

A abordagem tradicional da métrica *degree centrality* considera que os nós com mais ligações tendem a ser mais poderosos, visto que podem influenciar de forma direta outros elementos. No entanto, o facto de terem o mesmo grau não os torna igualmente importantes. Com base nesta constatação, Bonacich (1987) apresentou uma modificação da abordagem *degree centrality*, propondo a métrica de centralidade *eigenvector*, assente em valores próprios e vetores próprios de matrizes. A centralidade *eigenvector* (C_E) atribui a cada nó um valor proporcional à soma do valores dos seus vizinhos (Bonacich, 2007; Burguillo, 2018; Golbeck, 2015; Koutra e Faloutsos, 2018; Qasem et al., 2017). Esta métrica tem como objetivo medir a importância de um elemento da rede, em função da importância dos seus vizinhos. Ou seja, se os vizinhos de um nó forem importantes, mesmo que o nó tenha poucas ligações, logo um valor baixo de *degree centrality*, adquire uma centralidade *eigenvector* elevada.

Bonacich (1987) considera a métrica apropriada para medir o poder de um nó, formalizando-a da seguinte forma:

$$C_E(\alpha, \beta) = \alpha(I - \beta R)^{-1} R1 \quad (4)$$

Onde o parâmetro β permite alterar o grau e o sentido (positivo ou negativo) do peso do valor de cada unidade no valor de outras unidades. O parâmetro α é um vetor de escala, definido para normalizar o resultado, 1 é um vetor de coluna, I é a matriz de identidade e R é a matriz de adjacência[24]. É uma matriz quadrada de ordem n x n (n linhas e n colunas), onde a cada nó corresponde uma linha n_i e uma coluna n_j. As entradas da matriz, R_{ji}, registam pares de nós adjacentes, ou seja, se o nó n_i e n_j são adjacentes, então R_{ji} assume valor 1, caso contrário R_{ji} = 0 (Wasserman e Faust, 1994).

Para além das métricas de centralidade referidas salientam-se outras duas métricas também muito utilizadas e importantes para a análise de RSO. São elas a métrica *modularity class* e o *PageRank*. A primeira permite identificar subgrupos ou comunidades dentro de uma rede (Brandes e Erlebach, 2005; Newman, 2006). A literatura (Cherven, 2015; Moreira et al., 2019; Ozyer et al., 2013; Savic et al., 2019), mostra que a *modularity class* é amplamente utilizada como uma medida de qualidade de um subgrupo. Um elevado valor de *modularity class* indica uma estrutura de rede complexa. Essa estrutura descreve como a rede é subdividida em várias sub-redes que normalmente são designadas de comunidades.

A segunda, o algoritmo *PageRank*, inicialmente proposto por Brins e Page (1998) e Page et al. (1999) para medir a reputação das páginas *web*, tem sido muito estudada e utilizada durante as últimas duas décadas. Aplicada no âmbito das RSO, de acordo com a literatura (Bonchi et al., 2011; Fu et al., 2017; Gliwa e Zygmunt, 2015; Koutra e Faloutsos, 2018; Torres-Moreno, 2014), o *PageRank* é uma variação da métrica *eigenvector* e mede a importância de um nó contabilizando a quantidade e qualidade das suas ligações. O trabalho de Rieder (2012), analisou a evolução da métrica e as suas definições bem como evidenciou a sua importância para a SNA.

Como referido em Freire et al. (2021), a utilização ou combinação das métricas da SNA permite explorar as RSO para apoio à decisão, pois fornecem um entendimento mais profundo de como as entidades da rede (utilizador, mensagem e conceito) interagem juntas. Para além disso, permitem evidenciar propriedades e características das interações das RSO tais como importância, poder, prestígio, influência e difusão de informações dentro de uma rede (Arif, 2015; Izquierdo e Hanneman, 2006; Ruas et al., 2019; Saint-Charles e Mongeau, 2018; Wasserman e Faust, 1994). As métricas de centralidade, resumidas na Tabela 1,

[24] A matriz de adjacência permite definir, através de zeros e uns, a existência ou não de ligações entre pares de nós de uma rede (Hanneman e Riddle, 2005; Savic et al., 2019).

utilizam-se de diferentes formas para identificar a atividade, o controlo do fluxo de informação, a importância, o poder, o prestígio (*status*, popularidade) e a proximidade:

- *Atividade* - Pode medir-se em termos de ligações de entrada e de saída, quantificando as interações recebidas e ou enviadas por um nó específico. Um nó com um nível alto de atividade denota que mantém um elevado número de interações. Em sentido oposto, os nós com baixa atividade são periféricos na rede e não são ativos, visto que estão isolados;
- *Controlo do fluxo de informação* - Representa a capacidade de o utilizador controlar a difusão da informação dentro da RSO. Essa difusão de informação está relacionada, por exemplo, a recomendações para orientar outros clientes para novos produtos e/ou serviços que sejam do seu interesse, aumentando assim as possibilidades de vendas adicionais;
- *Importância* - A importância de um nó está relacionada com a capacidade de este estabelecer relacionamentos com outros utilizadores e pode depender da importância de seus vizinhos. Por exemplo, um utilizador é importante ou influente se for seguido por outros utilizadores importantes ou influentes;
- *Poder* - Identifica quem está amplamente envolvido nas interações estabelecidas com outros utilizadores e quem é mais visível na rede. Os nós que possuem um maior número de ligações tendem a ser mais poderosos, porque podem influenciar diretamente mais nós;
- *Prestígio* - Está associado ao nível de expansão e popularidade de um utilizador. Um utilizador com um elevado nível de prestígio tem contacto direto com muitos outros. Em sentido oposto, utilizadores com baixo prestígio estão isolados. A popularidade está relacionada com as interações criadas com os outros (por exemplo seguidores) e aos relacionamentos criados entre eles;
- *Proximidade* - descreve a proximidade de um nó face a outro e à rapidez com que um utilizador pode alcançar todos os outros na rede. Acompanhar esse tipo de utilizador é uma parte essencial num processo de negócio, pois eles conseguem alcançar rapidamente os seus clientes.

Tabela 1 - SNA: resumo da interpretação das métricas de centralidade

Métricas SNA	Interpretação das métricas de centralidade	Atividade	Controlo do fluxo de informação	Importância	Influência	Poder	Prestígio	Proximidade
Degree	Número de ligações que um nó possui. Nós com muitos vizinhos são centrais.	●	●	●	●	●	●	
In-degree	Número de ligações que apontam para um nó.	●	●	●	●		●	
Out-degree	Número de ligações com origem num nó.	●	●		●	●		
Closeness	Distância ou proximidade de um nó relativamente aos outros. Quão perto um nó está de todos os outros nós.	●			●	●		●
Betweenness	Número de vezes que um nó atua como uma ponte no caminho mais curto entre dois outros nós.		●		●			
Eigenvector	Nós ligados a nós centrais são eles próprios centrais.			●			●	●
PageRank	O nó é importante de acordo com a quantidade e a qualidade dos *links* que apontam para ele.			●			●	●
Modularity class	Identifica subgrupos ou *clusters*, enquanto indica a densidade da estrutura da rede.			●			●	●

3.3 Métodos de representação da SNA

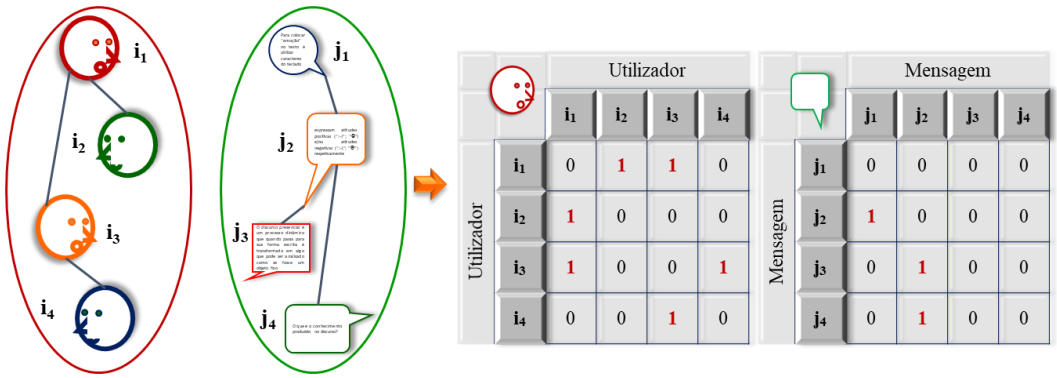
A SNA utiliza métodos formais, matrizes e grafos, para representar as RSO atendendo a três motivos fundamentais. O primeiro motivo é representar os padrões das interações sociais das RSO de forma compacta e sistematizada (Benjamin et al., 2017; Hanneman e Riddle, 2005; Koutra e Faloutsos, 2018; Liu et al., 2018; Pinheiro, 2011). O segundo prende-se com o facto das matrizes e dos grafos poderem ser modelados e formalizados pelos SI e TI, o que permite armazenar e transformar a informação mais rapidamente e com maior precisão (Hanneman e Riddle, 2005; Koutra e Faloutsos, 2018; Pinheiro, 2011; Rahman, 2017). Por último, as técnicas matemáticas e de visualização gráfica têm regras e convenções próprias que ajudam a comunicar de forma clara, identificando, nos dados, situações que não seriam identificáveis se fossem descritas apenas por palavras (Hanneman e Riddle, 2005). Quer o cálculo matemático, quer a visualização das RSO, pode ser feito com apoio de *software*, ele também responsável pelo desenvolvimento dos estudos de SNA. O *software* permite representar de uma forma simples, resumida e rápida um grande volume de informação, possibilitando sistematizar e descrever de forma exaustiva as relações sociais.

3.3.1 Técnicas matemáticas - matrizes

As técnicas matemáticas da SNA permitem modelar os nós e as ligações das RSO e muitas das suas características podem ser analisadas através da manipulação direta de matrizes (Benjamin et al., 2017; Koutra e Faloutsos, 2018; Scott, 2000). Habitualmente, os grafos são formalizados por uma matriz, com uma estrutura $n \times n$, onde n são as linhas e as colunas que representam os nós e as posições (i, j) descrevem a relação existente entre estes (Wasserman e Faust, 1994; Zheng e Skillicorn, 2017). Na representação matemática das RSO normalmente são utilizados dois tipos de matrizes: as matrizes de adjacência, cujas linhas e colunas representam o mesmo conjunto de nós, e as matrizes de afiliação, cujas linhas e colunas representam conjuntos distintos de nós.

Por definição, a matriz de adjacência permite definir, através de zeros e uns, a existência ou não de ligações entre pares de nós de uma rede (Hanneman e Riddle, 2005; Rahman, 2017; Savic et al., 2019; Wasserman e Faust, 1994). Na Figura 12, as tabelas apresentadas nas alínea (c) e (d) representam, respetivamente, as matrizes de adjacência do conjunto de nós das entidades utilizador e mensagem, apresentadas nas alínea (a) e (b).

Figura 12 - Rede de utilizadores e mensagens, e matrizes de adjacência correspondentes

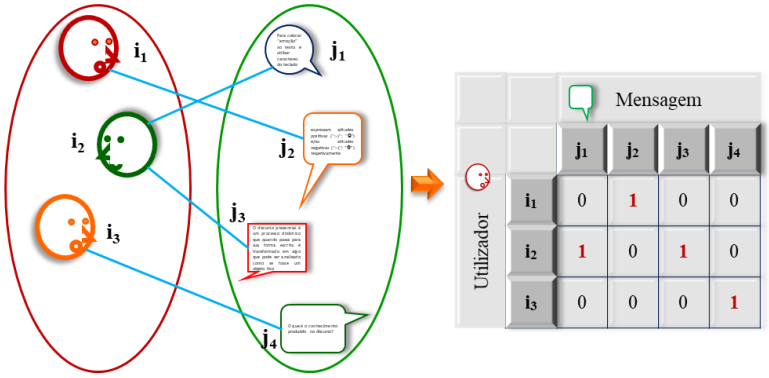


(a) Rede entidade Utilizador (b) Rede entidade Mensagem c) Matriz de adjacência entidade Utilizador d) Matriz de adjacência entidade Mensagem

Por definição, na matriz de afiliação as linhas representam os nós e as colunas representam eventos (Everett e Borgatti, 2005; Faust, 2005), e um 1 na linha *i* coluna *j* indica que o nó *i* participou no evento *j*. De acordo com Borgatti e Halgin (2011), para a SNA, o termo afiliação parte do pressuposto de que a afiliação num grupo ou participação num evento tem implícito um vínculo social. A Figura 13 ilustra essa definição. Neste caso, considera-se que os nós são utilizadores e os eventos são as mensagens trocadas por eles. A matriz (c) da Figura 13 exemplifica que utilizador i_2 (representado na figura em (a))

participou na troca discursiva com 2 mensagens (j_2 e j_3, representadas na figura em (b)) e os utilizadores i_1 e i_3 participaram, cada um, com 1 (j_1 e j_4 respetivamente).

Figura 13 - Rede de utilizadores e mensagens, e matriz de afiliação correspondente

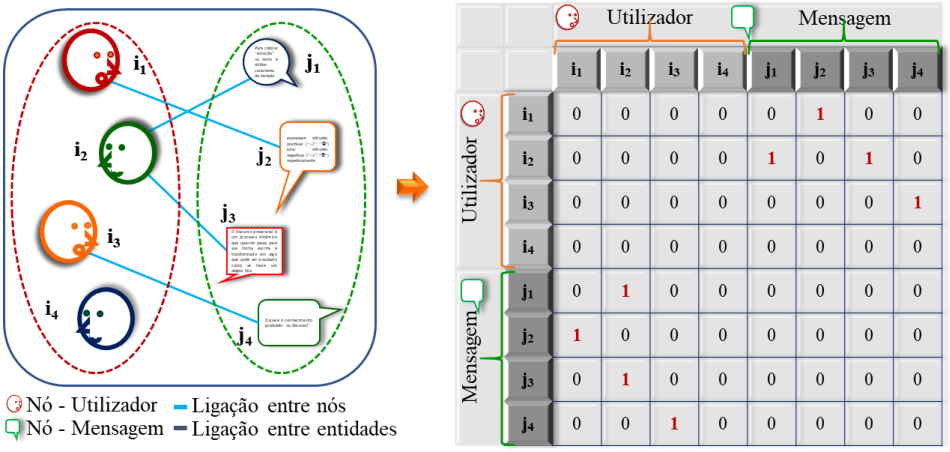


(a) Rede entidade Utilizador (b) Rede entidade Mensagem (c) Matriz afiliação

A maioria das redes são definidas por uma matriz de adjacência, onde os nós pertencem todos à mesma entidade, sendo designadas de redes *one-mode*. No entanto, quando a rede é composta por várias entidades (conjuntos distintos) estamos perante redes *two-mode* (Banerjee et al., 2017; Borgatti, 2009; Borgatti e Halgin, 2011; Carrington et al., 2005; Ikematsu e Murata, 2013; Opsahl, 2013). Nas redes *two-mode*, os nós de duas entidades distintas (ou mais) podem representar-se de duas formas diferentes: *affiliation graphs* e *bipartite graphs.*

Por definição, a *bipartite graph* representa uma rede em que os nós estão divididos em duas (ou mais) entidades e as únicas ligações entre elas situam-se entre os nós das diferentes entidades (Koutra e Faloutsos, 2018; Rahman, 2017; Vicario et al., 2017; Zheng e Skillicorn, 2017). Apesar de existirem dois tipos de representações, para Faust (2005) e Borgatti e Halgin (2011), formalmente, os *affiliation graphs* podem ser definidos como um *bipartite graph*. Ou seja, através de uma matriz de adjacência, onde só existem ligações entre nós dos dois conjuntos e não há arestas entre os nós que pertencem ao mesmo conjunto. A Figura 14 ilustra a representação do *bipartite graph* na sua matriz de adjacência correspondente. Neste caso, todas as ligações da *bipartite graph* da figura (a) são representadas na matriz (b), e situam-se entre o conjunto dos utilizadores e o conjunto das mensagens e vice-versa.

Figura 14 - *Bipartite graph* e a matriz de adjacência correspondente



(a) *Bipartite graph* (b) Representação matricial da *bipartite graph*

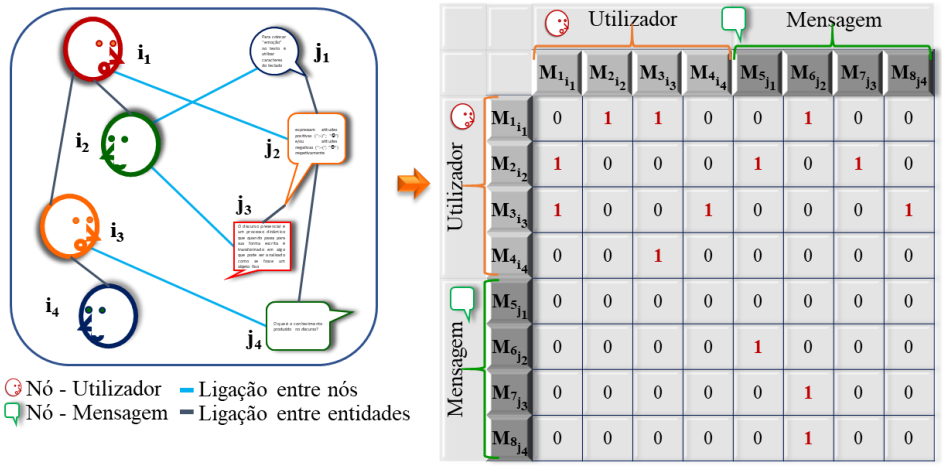
O princípio subjacente às redes sociais é simples. Uma rede social é um conjunto de indivíduos (nós) que podem ter relacionamentos (ligações) uns com os outros (Hanneman e Riddle, 2005). Em particular, o discurso *web* das RSO é constituído por três entidades (conjuntos distintos) e múltiplas relações entre elas, que podem ser representadas por um conjunto de matrizes da mesma dimensão, com os nós a ocupar a mesma posição em cada submatriz.

Nas RSO, as ligações entre utilizadores caracterizam-se por serem relações de interação. Quando os utilizadores trocam mensagens entre si, estabelecem uma ligação caracterizada como ação social ou ato discursivo. Os dados que as RSO produzem, podem assim modelar-se com técnicas matemáticas da SNA e formalizarem-se por uma matriz quadrada com atributos e variáveis da rede. As colunas e as linhas da matriz são os utilizadores, mensagens e conceitos e cada posição da matriz (encontro entre linha e coluna) descreve as ligações existentes entre eles. Segundo Everett e Borgatti (2005), este tipo de distinção é interessante para os estudos de rede, quando se pode supor que dois indivíduos (nó) que participaram do mesmo evento (nó) têm ou podem ter algum vínculo social entre eles.

Por exemplo, considerando a Figura 15 (b), que representa uma matriz *two-mode* com as entidades utilizador e mensagem, o valor 1 na linha *i* coluna *j* indica que um utilizador *i* publicou *online* uma mensagem *j*. Na tabela, o utilizador i_1 publicou *online* a mensagem j_2, e o utilizador i_3 publicou *online* a mensagem j_4. A Figura 15 ilustra como cada utilizador (nó) contribuiu com mensagens (nó) estabelecendo uma troca discursiva que representa uma ligação entre eles. Com a formalização e representação do discurso *web* é possível verificar

quem escreveu uma mensagem e a quem. A representação, quer visual, quer da matriz, permite verificar que o utilizador i_4 foi o único que não participou da troca discursiva e que i_2 foi o mais participativo.

Figura 15 - Rede *two-mode* e a matriz de afiliação correspondente



(a) Rede *two-mode*

(b) Matriz *two-mode*

Na matriz, o utilizador i_2 tem ligação com o utilizador i_1 porque participou numa troca discursiva contribuindo com a mensagem j_1. O utilizador i_1 tem ligação com o utilizador i_2 porque respondeu com a mensagem j_2. Neste exemplo, a matriz representa duas entidades (i, j) do discurso *web*, respetivamente utilizador e mensagem.

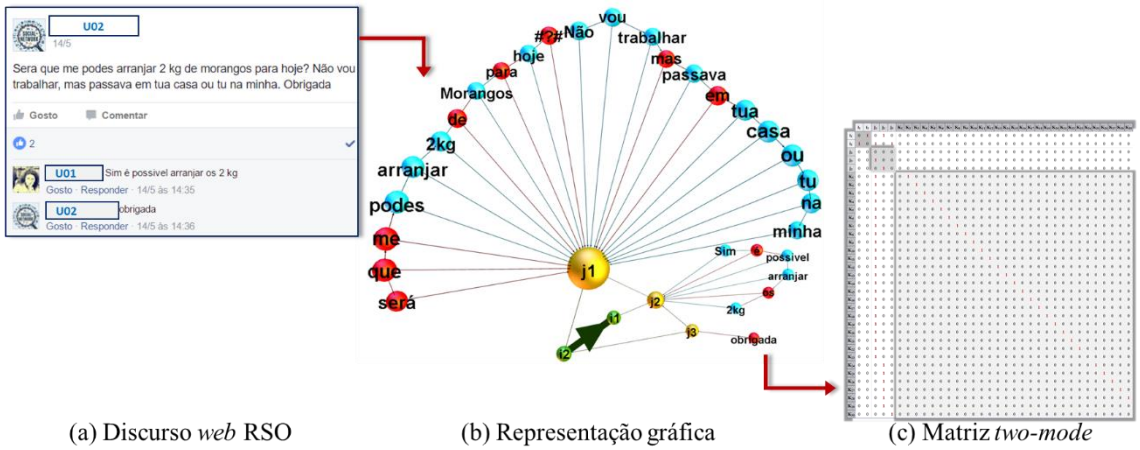
3.3.2 Métodos descritivos - visualização gráfica

A visualização das redes é uma componente da SNA (Benjamin et al., 2017; Freeman, 2005; Koutra e Faloutsos, 2018; Rahman, 2017) que simplifica a compreensão e interpretação da informação, tornando o processo de extração de conhecimento mais simples. Para Cherven (2013), à medida que se começaram a criar exemplos gráficos que mostram as redes interligadas do mundo das RSO, os grafos passaram a ser uma componente essencial no mundo da visualização. Segundo Panda et al. (2014), nas últimas décadas, os investigadores da área da SNA desenvolveram várias técnicas sofisticadas para analisar e visualizar dados de rede. As representações gráficas utilizam-se para representar as interações entre os indivíduos das RSO e são essenciais para estudar como é que a informação flui (Alhajj e Rokne, 2018; Borgatti et al., 2014; Diestel, 2017). A SNA suporta-se da teoria dos grafos para representar matematicamente grupos *online* de indivíduos com interações entre si. Perante grandes volumes de dados, as estruturas e a

composição das RSO podem ser difíceis de interpretar só com matrizes e a sua representação gráfica auxilia todo o processo de análise.

Por exemplo, como ilustra a Figura 16, as interações entre dois utilizadores apresentadas em (a) são mais fáceis de interpretar através da visualização do grafo (b), comparativamente com as matrizes de adjacência (c). Com apoio da SNA, o discurso *web* é estruturado e os dados são encadeados de forma coerente para facilitar o seu acesso e a sua análise. Através da visualização do grafo apresentado em (b), é possível identificar como cada interveniente se relaciona com os outros, como contribuiu com informação, e resumir o conteúdo das mensagens.

Figura 16 - Exemplo de representação matricial e gráfica do discurso *web*

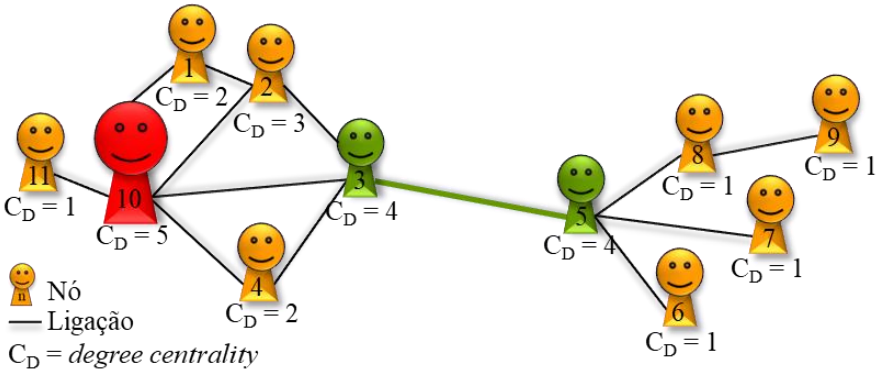


(a) Discurso *web* RSO (b) Representação gráfica (c) Matriz *two-mode* (matriz de adjacência)

A visualização gráfica da informação é muito utilizada como ferramenta para interpretar dados, isto é, descobrir padrões nos dados, ligações e a sua estrutura (Isson, 2018; Krempel, 2011; Lima, 2011; Nisbet et al., 2018). A visualização e exploração dos dados permite interpretá-los, e efetuar o que a literatura referente à análise de dados (Bassan e Sarkar, 2014; Moreira et al., 2019; Provost e Fawcett, 2013) designa de *data understanding*. Entender os dados permite identificar e perceber a relação que existe entre as diferentes fontes de dados e os seus atributos-chave e identificar se são necessárias algumas transformações. Para Kantardzic (2011), esse entendimento é feito utilizando técnicas estatísticas e de visualização, para avaliar a qualidade e descrever os dados. De acordo com Provost e Fawcett (2013), significa perceber quais os pontos fortes e quais as limitações dos dados, pois muito raramente há uma correspondência exata entre um problema e a informação correspondente.

No contexto da análise da estrutura da rede de utilizadores das RSO, a visualização gráfica justifica-se por dois motivos. Em primeiro lugar para identificar nós importantes, visto que as diferentes definições e métricas de centralidade da SNA capturam de forma diferente a origem do poder na rede, o que resulta em interpretações muito diferentes das estruturas sociais. Considere-se, por exemplo, a Figura 17 que representa uma pequena RSO em que cada ligação entre os seus elementos traduz a métrica *degree centrality*.

Figura 17 - Representação gráfica de utilizadores-chave



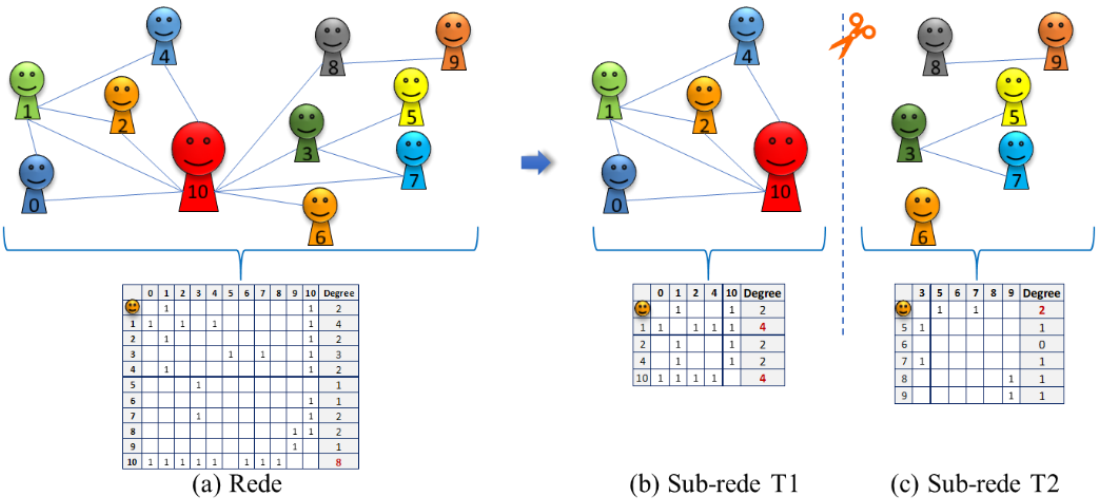
Nesta rede e de acordo com a definição de centralidade da métrica *degree centrality*, observa-se que o utilizador 10 é o mais central visto que tem o valor mais elevado (C_D=5). Todavia, pela visualização da rede, observa-se que o par de nós 3 e 5 chega a mais utilizadores do que o 10. Isto significa que a ligação entre eles (3 e 5) é importante e, se deixar de existir, a rede divide-se em duas sub-redes isoladas. Nessa perspetiva, os nós 3 e 5 juntos são mais importantes para a rede do que o nó 10. Como referem Laumann et al. (1992), é preciso ter cuidado para não omitir, por descuido ou utilização meramente conveniente dos dados, um indivíduo muito interveniente dentro de uma rede, visto que, de acordo com as definições de centralidade, são considerados utilizadores-chave ou pontes. A interpretação gráfica da rede permite identificar eventuais conjuntos de utilizadores-chave e ter em atenção estes casos.

Em segundo lugar, a análise e exploração visual é útil para observar as características dinâmicas das RSO, visto que a variável tempo tem impacto na continuidade dos dados. Como refere Aggarwal (2011), na grande maioria das redes de interação contínua, as mensagens trocam-se a uma taxa muito elevada e podem ocorrer grandes fluxos de informação na rede. Esta continuidade está associada à perda de ligações entre os nós, o que provoca transformações na rede. As RSO são constituídas por padrões de relacionamentos sociais e é necessário identificar pedaços de tempo que capturem a natureza e as

características que interligam as interações às respetivas trocas discursivas. Nesse sentido, é necessário definir de forma cautelosa o que a rede representa e os seus limites. Para visualizar os dados de forma contínua, capturam-se vários *snapshots*[25] que, quando representados graficamente, mostram resultados relacionados e delimitados por um intervalo de tempo definido (por exemplo, intervalos de um dia, uma semana ou um mês). De acordo com Carley (2015) e Zaidi et al. (2018) os *snapshots* são a técnica mais utilizada para observar e descrever alterações em redes dinâmicas.

Por exemplo, considerando a Figura 18, ao dividir-se uma rede em duas sub-redes, T1 e T2, em que cada uma é um intervalo de tempo, perdem-se as interações que ocorrem entre o momento T1 e o momento T2. Cada momento (T1 e T2) define um intervalo fechado à direita e à esquerda. Estes intervalos, como refere Aggarwal (2011), são estáticos. As interações entre os utilizadores do intervalo T1, ligadas às interações do intervalo T2, isto é, interações em curso, estendem-se para além dos limites do intervalo. As interações dos nós 3, 6, 8 e 10 não estão incluídas nem na sub-rede T1, nem na sub-rede T2. Recorrendo à visualização gráfica é possível identificar as pontas soltas, ou seja, ligações perdidas.

Figura 18 - Impacto do intervalo de tempo



Uma das consequências da utilização de *snapshots* é que alguns nós ficam com pontas soltas e, como referem Zaidi et al. (2018), isso pode levar a perdas de informação significativa. Isto acontece porque as ligações entre os nós vão para além da sua ocorrência,

[25] S*napshot*s são amostras estáticas (Aggarwal, 2011) utilizadas para recolhas de dados temporais em intervalos de tempo, isto é, são cortes transversais da rede. A análise tem o seu foco nas alterações de uma rede para a outra, sem qualquer referência (explícita) para a sequência de alterações que provocam a mudança de uma para outra rede. Nas bases de dados os *snapshot* representam os objetos temporais (Jensen et al., 1996).

visto que estão limitados pela dimensão do intervalo de tempo. Como refere Lima (2011), o tempo é uma das variáveis mais difíceis de formalizar e em particular um *snapshot* pode dar muito pouca informação sobre uma RSO. Os conjuntos de dados dos intervalos de tempo são visões parciais da rede com pares de nós desligados uns dos outros e um número insignificante (ou não) de ligações, entre utilizadores, perdidas ou não consideradas. A visualização gráfica, de sequências de períodos temporais, auxilia na identificação das alterações que ocorrem na rede e permite avaliar se as ligações entre os nós são ou não significativas, bem como se existe ou não perda de informação.

3.3.3 Ferramentas de extração de dados e visualização gráfica

Os dados das RSO podem recolher-se de várias formas (Jamali e Abolhassani, 2006; Marsden, 2005; Van Duijn e Vermunt, 2006). Abordagens, ainda muito utilizadas dentro das ciências sociais, englobam os inquéritos, as entrevistas, a observação e as fontes secundárias, que utilizam métodos próprios para recolha de dados (Leech et al., 2018). Alguns estudos (Mesa et al., 2018; Rapp et al., 2017; Sadovykh e Sundaram, 2016, 2017) utilizam a netnografia, desenvolvida por Kozinets (2010) como técnica de pesquisa de *marketing online* para estudar as percepções do consumidor. Outros estudos (Erétéo et al., 2011; Himelboim et al., 2013; Kefi et al., 2016; Réda et al., 2017; Velde et al., 2015) utilizam abordagens manuais de codificação de dados recorrendo a equipas de pessoas.

Porém, existe uma fonte de dados que não é apenas grande, diversificada e dinâmica, mas é de acesso livre, a *web* social (Mika, 2007; Russell e Klassen, 2019). Como referido em Freire et al. (2021), as RSO tornaram possível a recolha de dados referentes às opiniões e experiências de um amplo conjunto de pessoas, sem necessidade de inquéritos formais, usualmente presenciais.

Enquanto que nos métodos tradicionais de recolha de dados, tais como inquéritos ou questionários, não existe grande complexidade na preparação dos dados, em sentido oposto, os dados das interações sociais têm de passar por várias fases até à sua utilização final. Os processos, recursos de SI e de TI, técnicas, metodologias, de que dependem essas fases é que fazem com que a análise de RSO seja uma área complexa. Para um utilizador comum nem sempre é fácil encontrar instruções técnicas e práticas de como utilizar

Application Programming Interfaces (APIs[26]), os *softwares* de extração de dados e como os encadear dentro do processo de análise de RSO. Para além disso, a investigação prática ainda está muito concentrada em artigos e livros teóricos, que abrangem apenas os aspetos metodológicos e matemáticos da análise de RSO. Isso acontece porque estamos perante domínios diferentes de investigação, que abordam de forma distinta cada situação.

Atualmente não é complicado extrair dados das RSO, pois nestes últimos anos para além de estar disponível uma grande variedade de *software*, têm sido desenvolvidas ferramentas *user-friendly* para esse efeito. Para Moreira et al. (2019), o que alterou o cenário foi que a recente evolução tecnológica nas áreas do processamento, armazenamento e transmissão de dados, associada a *softwares* mais avançados, provocaram uma redução dos custos e aumentaram a capacidade dos SI e TI. As próprias plataformas de RSO (*Facebook*, *Twitter*, etc.) disponibilizam interfaces, APIs, ou serviços para extrair os seus dados.

As ferramentas de extração de dados podem ser implementadas utilizando várias possibilidades de linguagens de programação (por exemplo, *R*[27], *Python*[28], *Java*[29], entre outras), que devolvem *outputs* de dados em vários formatos (por exemplo, GDF[30], TSV[31], CSV[32], XML e JSON[33]). A Tabela 2 identifica algum *software* e algumas ferramentas para extração de dados do *Facebook*. De todo o *software* de extração de dados analisado, os mais *user-friendly* e por isso os que têm sido mais utilizados pela comunidade académica para investigação são o *NodeXL*[34] e o *Netvizz*[35].

---

[26] API é um conjunto de rotinas e padrões estabelecidos por um *software* para a utilização das suas funcionalidades por outros *softwares*.

[27] *R* é um ambiente computacional e uma linguagem de programação para manipulação, análise e visualização gráfica de dados (Torgo, 2017). Acesso em 2 de dezembro de 2019, disponível no portal do *The R Project for Statistical Computing*: https://www.r-project.org/.

[28] *Python* é uma linguagem de programação de alto nível, *open-source*, para criação, manipulação e estudo de estruturas dinâmica (Al-Taie e Kadry, 2017).

[29] *Java* é um ambiente computacional e uma linguagem de programação (Davis, 2019).

[30] O GDF é o formato de ficheiros utilizado pela *GUESS graph analysis toolkit*, onde os dados estão separados por tabulações ou vírgulas, podendo ser abertos com outros softwares.

[31] O formato *Tab-Separated Values* (TSV) guarda os dados num ficheiro em formato tabular (colunas).

[32] O *Comma-separated values* (CSV) é um formato de ficheiro com informação em texto simples (caracteres alfanuméricos) separada por vírgulas e permite que a mesma seja guardada em tabelas estruturada.

[33] JSON é um formato de texto embutido no *JavaScript* (Davis, 2019).

[34] Acesso em 8 de novembro de 2014, disponível no portal do *NodeXL*: http://nodexl.codeplex.com/.

[35] API descontinuada e sem acesso ao público desde 4 de setembro de 2019. Acesso em 16 de dezembro de 2019, disponível no portal do *The Politics of Systems*: http://thepoliticsofsystems.net/2015/01/the-end-of-netvizz/.

Tabela 2 - *Software* e ferramentas para extração de dados do *Facebook*

Conhecimento Programação	*Software*/Ferramenta	*Link*	Características
Não requerido	*Digitalfootprints*	http://digitalfootprints.dk/footprints.dk/	Acesso restrito com *login* e licença
	Discovertext	https://discovertext.com/	
	Infoextractor	http://www.infoextractor.org/	*Open source*
	Facebook Graph API	https://developers.*Facebook*.com/docs/graph-api/overview	*Open source*
	NodeXL	https://www.smrfoundation.org/nodexl/	*Shareware*
	Nvivo/Ncapture	https://www.qsrinternational.com/nvivo/home	*Open source*
Requerido	*Facebook Python SDK*	http://*Facebook*-sdk.readthedocs.io	*Open source*
	Facepager	https://github.com/strohne/Facepager	*Open source*
	Pattern	http://www.clips.ua.ac.be/pattern	*Open source*
	RFacebook	https://CRAN.R-project.org/package=R*Facebook*	*Open source*
	SocialMediaMineR	https://CRAN.R-project.org/package=SocialMediaMineR	*Open source*

O *NodeXL* é um *add-in* para o *Microsoft Excel* que, para além de permitir exportar dados, também permite explorar e visualizar RSO (Golbeck, 2015). Esta ferramenta tem alguma autonomia em todo o processo que compõe o fluxo de análise de RSO que se inicia com a recolha dos dados da rede e passa por várias etapas até à visualização e criação de relatórios finais. O *NodeXL* permite que não-programadores calculem métricas da SNA e visualizem de forma rápida uma rede com o *Excel* (Hansen et al., 2011). Esta ferramenta é, no entanto, restritiva e limitativa quando se pretende analisar, em simultâneo, as três entidades de análise do discurso *web*, utilizador, mensagem e conceito (*user*, *post*, *concept*) e construir redes que não sejam as que a própria ferramenta recolheu.

O *Netvizz* era uma API do *Facebook* que podia ser acedida digitando o seu nome na caixa de pesquisa do *Facebook*. De acordo com Rieder (2013), a ferramenta *Netvizz*, foi desenvolvida para ajudar os investigadores na extração de dados do *Facebook*. A ferramenta extraía dados em "bruto" da RSO *Facebook* e, para além da estrutura da rede, também extraía em formato tabular o conteúdo (texto) quer das mensagens quer dos comentários a eles associados. As fontes de dados extraídas, quando relacionadas entre si, permitiam estruturar as trocas discursivas produzidas pelos utilizadores.

A SNA também está estreitamente relacionada com a existência de *software* próprio para representar e visualizar a informação assente em matrizes e grafos, ou seja, criar imagens das RSO. O trabalho de Huisman e Van Duijn (2011) apresenta uma visão geral de 56 *softwares* e ferramentas para SNA. Muitas dessas ferramentas requerem conhecimentos de programação e terminologia de rede mais técnica, tornando-se um desafio para aqueles que não detêm essas competências para importar e explorar o sentido dos dados de RSO.

Todavia, nesta última década, todo o processo de visualização de dados sociais tem sido simplificado com a evolução tecnológica e a utilização dos computadores pessoais.

Para além disso, pela literatura (Chen et al., 2012; de Nooy et al., 2018; Hansen et al., 2011; Krishna et al., 2018; Lima, 2011; Ravindran e Garg, 2015; Smith et al., 2009b; Torgo, 2017), percebe-se uma tendência crescente na disponibilização de s*oftware* e linguagens de programação (com vasta documentação), com interface gráfica e compatíveis para importação/exportação de dados com outros *softwares*. O *software* de visualização de redes utiliza *input* de dados estruturados em diferentes formatos (GDF, TSV, CSV, etc.) onde todas as interações entre utilizadores e alguns metadados são definidos e codificados.

As ferramentas para análise e visualização de RSO assentam na teoria dos grafos, e têm emergido de muitos grupos de investigação. Dentro destas ferramentas destacam-se o *Gephi* (Bastian et al., 2009) e a sua documentação (Cherven, 2013, 2015) bem como o *plugin NodeXL* para *Excel* (Smith et al., 2009a) e literatura associada (Hansen et al., 2011). Estas duas ferramentas, muito recentes e de fácil utilização, permitem sistematizar e resumir graficamente um grande volume de informação que descreve de forma exaustiva as relações sociais. A visualização através dos grafos, por si só, oferece uma perspetiva diferente de análise que só é possível com a ajuda de *software* de SNA.

O *NodeXL* contém um vasto conjunto de funcionalidades que permitem calcular métricas, filtrar vértices e respetivas interações, personalizar o aspeto e o *layout* dos grafos (Hansen et al., 2011), entre outras. Esta ferramenta, para identificar grupos na rede, utiliza os algoritmos tradicionais de deteção de comunidades em sistemas complexos. De acordo com Smith et al. (2009a), a utilização de ferramentas de visualização ajuda a identificar novas informações, especialmente quando as propriedades visuais de um nó, tais como o tamanho ou a cor, são combinadas com métricas da SNA (por exemplo, de centralidade).

Apesar da existência de outras ferramentas de visualização, o *Gephi* tem sido o mais utilizado pelos investigadores (Heymann e Le-Grand, 2013; Kollwitz et al., 2018). Caracteriza-se por ser um *software open-source* para exploração e manipulação de redes, desenvolvido para importar, visualizar, filtrar, manipular e exportar todo o tipo de redes (Bastian et al., 2009). De acordo com Bastian et al. (2009), o *Gephi* apresenta um conjunto de características que permitem realizar uma análise de rede através de processos analíticos/estatísticos e de técnicas de análise exploratória dos dados.

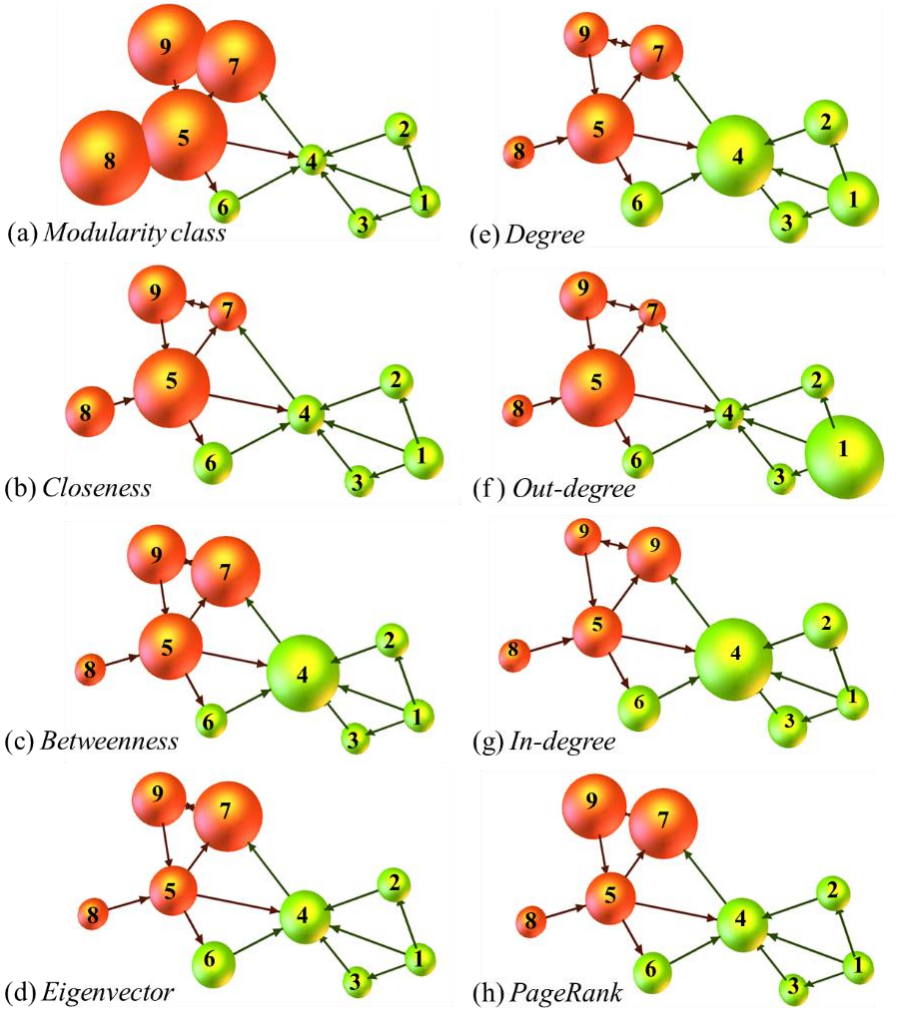
Como refere Freeman (2005), com a utilização de *software* padrão é possível produzir automaticamente imagens que identificam e chamam a atenção para subconjuntos específicos de nós, através da atribuição de símbolos ou cores distintas para identificá-los. De acordo com o autor, estas características são reveladas olhando simplesmente para as imagens. Golbeck (2015) corrobora essa afirmação quando refere que a SNA permite criar imagens e fazer cálculos que caracterizam o desempenho dos nós dentro da rede.

O *software Gephi*, em particular, permite também aplicar automaticamente classificações numéricas para a aparência visual das redes em termos de cores e tamanho dos nós. A Figura 19 ilustra oito perspetivas da mesma rede, após aplicação das métricas mais utilizadas em SNA. Como se pode observar, em cada *layout* o tamanho dos nós altera-se consoante as características específicas da métrica representada graficamente. Por exemplo, no *layout* (b) da Figura 19, os nós de maior dimensão correspondem a elementos da rede com maior *closeness centrality*. Já no *layout* (e), os nós maiores correspondem a elementos da rede com maior *degree centrality*. Assim, um nó maior representa, naturalmente, um valor maior da métrica. A visualização dos *layouts* da rede, permite várias interpretações.

Em cada *layout* da Figura 19, de acordo com a métrica especificada, os nós de maior dimensão são os mais centrais e os nós mais pequenos são os menos centrais. Numa primeira análise à figura, observa-se que existem grandes diferenças entre os oito *layouts* que representam a mesma rede. O *layout* (a) representa a métrica *modularity class*, que permite identificar subgrupos onde cada cor representa um grupo, evidenciando a densidade da estrutura da rede. Através da representação gráfica da métrica *modularity class*, identificam-se dois grupos distintos na rede: um representado pela cor verde e outro pela cor laranja.

Num trabalho de investigação vale a pena analisar qualquer elemento que tenha um valor elevado de centralidade, independentemente do que cada métrica represente. Por exemplo, considerando o nó 4 da rede, observa-se no *layout* (e) que o mesmo tem um elevado *degree centrality*. Em termos gerais pode referir-se que é o elemento mais ativo e com mais ligações da rede. Todavia, quando analisados os *layouts* (f) e (g), *out-degree* e *in-degree*, respetivamente, verifica-se que existem mais ligações com origem nos outros nós, do que dele para os outros. As métricas *in-degree* e *out-degree*, são apropriadas para representar o fluxo de informação da rede. Quando aplicadas numa rede, a espessura das ligações entre os nós representa o volume do fluxo de interações entre eles.

Figura 19 - Visualização gráfica das métricas SNA



Pelo *layout* (b) observa-se que o nó 4 da rede tem um baixo valor de *closeness centrality* e, pelo *layout* (c), que tem um valor elevado de *betweenness centrality*. Aplicada aos diferentes tamanhos dos nós, a métrica *closeness centrality* representa independência e reflete as diferenças na participação. A métrica evidencia quem comunica com outros utilizadores da rede, através de um número mínimo de intermediários e, em termos gerais, pode-se referir que o nó tem pouco acesso a outros elementos da rede. Relativamente à métrica *betweenness*, infere-se que um nó que tem um valor elevado tem uma posição importante na rede, é influente e a sua eliminação pode cortar ligações a outros grupos.

As métricas *eigenvector* e *PageRank*, representadas nos *layouts* (d) e (h), identificam os nós mais centrais em termos da estrutura global da rede. Estas métricas indicam se um nó da rede é mais ou menos central relativamente a outros, considerando a distância entre todos os nós. Na análise de redes é importante considerar que a utilização de

várias métricas reforça a fiabilidade dos resultados. Cada métrica permite identificar e classificar diferentes características dos nós, quer em termos individuais quer globais.

3.4 Resumo

Para compreender os fenómenos que ocorrem nas RSO, um dos principais objetivos da SNA é capturar as características estruturais da rede, isto é, estudar o padrão estrutural das ligações entre os nós. Assim sendo, abordaram-se neste capítulo os conceitos-chave da SNA e descreveram-se as métricas mais relevantes quer da estrutura global da rede, quer de âmbito individual dos seus nós. Em traços gerais, a SNA assenta na utilização de métodos numéricos apropriados para a análise dos vínculos das entidades das RSO. Nestes métodos encontra-se embutida uma fundamentação matemática para modelar as interações sociais e muitas das características fundamentais das RSO podem ser analisadas através da manipulação direta de matrizes.

Neste capítulo sistematizaram-se os conceitos e as métricas de centralidade da SNA. Calcular a centralidade de um nó tem como objetivo identificar a posição deste, relativamente às trocas discursivas que faz e à comunicação que efetua dentro da rede. Este conceito traz consigo associada a ideia de poder, influência e prestígio. De forma resumida, a centralidade baseada na métrica *closeness centrality* mede o desempenho do fluxo de comunicação, visto que está relacionada com o tempo que uma informação demora a ser partilhada por todos os elementos da rede. A centralidade fundamentada na métrica *betweenness centrality* de um nó, destaca o controlo que o elemento tem na comunicação com os outros e a eficiência da rede. As métricas de *closeness centrality* e *betweenness centrality*, assentam no pressuposto de que a informação, ou conteúdo, é difundido unicamente ao longo dos possíveis caminhos mais curtos (distância geodésica). Por último, a métrica *degree centrality* mede a influência direta que um nó tem em relação aos seus contactos diretos. Esta métrica destaca a existência de nós importantes dentro da rede, ou seja, a atividade concentrada à volta de determinados elementos.

Foram também examinados os métodos de representação utilizados pela SNA. A estrutura da rede pode ser modelada através de grafos e, por sua vez, estes podem caracterizar-se a partir de várias métricas formalmente definidas e geralmente utilizadas nos estudos e resolução de problemas de redes. Nesse sentido, a SNA proporciona uma análise visual e matemática das relações humanas, das trocas de comunicação e auxilia nos processos de estruturação dos dados semânticos. É a partir da estrutura da rede de

utilizadores que é possível estabelecer uma relação entre o processo de comunicação e o respetivo discurso *web* como um todo. Por um lado, as matrizes de adjacência permitem representar formalmente as três entidades do discurso *web* produzido nas RSO através de rede *two-mode*. Por outro, a visualização de grafos permite uma interpretação mais intuitiva dos dados dos diferentes níveis de análise e ajuda a acelerar não apenas o seu processamento, mas também a sua análise semântica. A visualização da informação da rede em diferentes *layouts*, permite representar não só a estrutura de rede de utilizadores, mas também obter resumos das redes semânticas, destacando com métricas da SNA os conceitos mais representativos das mensagens. Pode considerar-se que a representação gráfica das entidades *utilizador*, *mensagem* e *conceito*, é um resumo do discurso produzido nas RSO.

4 Processos de decisão e RSO

A utilização da SNA para apoio à decisão ainda é uma área emergente de investigação (Póvoa et al., 2017) e a análise dos conteúdos da *web* social tem por isso muitas questões ainda em aberto. Todavia, as RSO articuladas quer com a SNA, quer com os SI e as TI ajudam nessa procura de informação, quer a nível individual quer a nível organizacional, alterando a forma como as decisões podem ser tomadas. Este facto favorece uma mudança de paradigma, na forma como a informação é adquirida, consultada, tratada e utilizada para o apoio à decisão numa qualquer situação.

4.1 Processos e modelos de apoio à decisão

Mintzberg et al. (1976) definem "decisão" como um compromisso específico para uma ação, e "processo de decisão" como um conjunto de ações e fatores dinâmicos que se inicia com a identificação do estímulo para ação e termina com um compromisso específico dessa ação. Para Hatamura (2006), a tomada de decisão é escolher entre várias opções, onde existem muitas alternativas possíveis, sendo necessário definir a probabilidade de uma opção para 1 (100%) e a das restantes para 0. O autor considera essencialmente o "fazer acontecer" para escolher uma opção entre muitas.

Noutra perspetiva, Niu et al. (2009) e Nutt e Wilson (2010) definem o conceito de decisão como sendo um processo cognitivo, que envolve diferentes tarefas cognitivas, tais como a recolha de informação, avaliação da situação, produção e seleção de alternativas e implementação de soluções. Rosenfeld e Kraus (2018) definem a tomada de decisão como o processo pelo qual um caminho de ação é escolhido, com antecedência e com um objetivo claro, a partir de um conjunto de alternativas. Esta visão coaduna-se com outras definições ou correntes teóricas e em particular com a de Simon (1977).

Na literatura tradicional é possível encontrar vários modelos e teorias referentes ao processo de decisão (Das e Teng, 1999; Huber, 1981; Mintzberg et al., 1976, entre muitos outros). Nomeadamente a obra de Nutt e Wilson (2010) descreveu e analisou diferentes perspetivas teóricas e terminou com uma discussão sobre a possibilidade de criar uma teoria unificada de tomada de decisão. Contudo, apesar dos inúmeros conceitos existentes, é possível observar que o modelo de Simon (1977) é o mais reconhecido e utilizado, por ser

acessível e de fácil visualização. Exemplo disso são os estudos, na área da tomada de decisão, apresentados em Bucciarelli et al. (2018) e inspirados no trabalho de Simon.

O modelo de decisão proposto por Simon (1977) encontra-se dividido em quatro fases, exibidas na Figura 20, com uma revisão sistemática entre elas e, segundo Turban et al. (2007), é o mais conciso na caracterização da tomada de decisão racional. De acordo com Huber (1981), o modelo racional retrata um ambiente em que as decisões organizacionais são consequência de unidades organizacionais, que utilizam a informação racional para efetuarem escolhas de forma intencional, em nome da organização.

Figura 20 - Modelo de decisão

Fonte: adaptado de Simon (1977)

Relativamente às fases do processo de decisão, de acordo com a literatura (Filip et al., 2017; Laudon e Laudon, 2011; Turban et al., 2007; Vercellis, 2011), podem descrever-se como: fase da Inteligência, que envolve a análise do ambiente, de forma intermitente ou contínua, e inclui várias atividades que procuram identificar situações, problema ou oportunidades; fase de Projeto, que implica encontrar ou desenvolver e analisar possíveis cursos de ação (segundo os autores, inclui a compreensão do problema e análise das alternativas disponíveis, com base na viabilidade destas); fase da Escolha, definida como o ato crítico da tomada de decisão, visto que é a única em que a decisão é executada e onde o compromisso de seguir um determinado curso de ação é feito. Os autores salientam ainda que a fronteira entre as fases de projeto e escolha não são sempre claras. No entanto, é possível passar de uma para outra, visto que podem ser necessárias alterações na fase de projeto ao escolher-se uma alternativa; por último, a fase de Implementação, onde a alternativa escolhida é implementada. De uma forma simplista Turban et al. (2007) referem que esta fase significa colocar uma solução recomendada, a funcionar. Contudo, os autores evidenciam que a implementação pode ser complicada, visto que pode ser um processo longo, que envolve fronteiras vagas.

A literatura mostra que os modelos clássicos de tomada de decisão estão associados à existência de processos racionais e sequenciais em contextos organizacionais tradicionais. As teorias e modelos de decisão tradicionais assentam na ideia de que perante um problema é necessário identificar alternativas para o resolver. Para as teorias clássicas, de um modo geral, o processo de decisão começa com a consciência de um problema de decisão e termina com uma solução escolhida, entre um número finito ou infinito de alternativas.

Contudo, o atual volume de informação proveniente da *web* social e a necessidade de decidir em tempo real, alteram a forma como as organizações se comportam. Isso tem implicações e consequências nos processos de tomada de decisão existentes e, como resultado, um efeito direto nos modelos de tomada de decisão. Por isso, a crescente importância atribuída pelas empresas à tomada de decisão em contexto *web* requer a definição e implementação de mecanismos mais eficientes para apoiar os processos de decisão. Como referido por Robinson et al. (2015), a capacidade de compreender e analisar dados altamente interligados é um fator chave para as empresas superarem os seus concorrentes.

A informação extraída das RSO para apoio à decisão em tempo real está em constante mudança, atualização e evolução. Este facto faz com que se inverta o paradigma na era dos dados e da informação. Isto é, perante uma situação ou um acontecimento, o(s) problema(s) pode(m) não estar à partida identificado(s) e a análise dos dados permite não só identificá-lo(s), mas também antecipá-lo(s) (vendo, por exemplo, tendências de mercado ou identificando perfis de utilizadores para direcionar produtos ou serviços).

O processo de decisão assente em dados das RSO pretende ser mais preditivo e analítico do que o processo de decisão tradicional com atuação *a posteriori*. O objetivo, como referem Moreira et al. (2019), não é prever o que vai acontecer, mas sim compreender o quão prováveis são os resultados de um determinado acontecimento. Também para Bahga e Madisetti (2019), a análise preditiva tem como finalidade perceber a frequência ou resultado de um acontecimento, para responder à questão "o que é provável que aconteça?". Nesse sentido, o conhecimento e valor extraído dos dados das RSO através da SNA, pode fornecer novo discernimento e algumas oportunidades que auxiliem as várias fases dos processos de tomada de decisão.

Apesar dos avanços tecnológicos e as vantagens que os SI e as TI trazem para o apoio à decisão organizacional, o processo de análise dos dados da *web* social ainda é algo complexo para algumas organizações. Este processo deixa qualquer decisor com um

conjunto de dados para transformar em informação, para fazer escolhas e/ou estudar opções no sentido de tomar melhores e mais rápidas decisões. Transformar rapidamente os dados em informação para apoio à decisão é, por isso, uma das dificuldades que os decisores enfrentam.

4.2 Apoio à decisão no contexto das RSO

A literatura mostra que as áreas do conhecimento, da decisão e dos SAD, estão interligadas quer com as RSO, quer com a SNA. Power et al. (2011) analisando o passado e o futuro dos SAD referem que a vasta disponibilidade de recursos computacionais e o aumento das competências tecnológicas dos utilizadores incentivam a mais investigação e desenvolvimento de ferramentas que auxiliem a eficácia da tomada de decisões complexas. Para os autores, a conjugação dos sistemas de decisão e das plataformas de RSO é uma nova ferramenta para apoio à decisão, quer ao nível dos sistemas de recomendação, quer dos sistemas de decisão colaborativa.

Alguns autores (Adam, 2008; Herring, 2013; Sueur et al., 2012) defendem que a SNA é particularmente adequada para o estudo da tomada de decisão, pois permite reconhecer a natureza dinâmica das redes, disponibilizando ferramentas e técnicas para medir e avaliar a mudança. Também Shum et al. (2011) evidenciam a ideia de que o relacionamento social e os conteúdos produzidos nas RSO podem ser analisados em conjunto para apoiar o processo de tomada de decisão. Como refere Marmo (2011), a SNA e a mineração *web* dão um grau de detalhe inovador na análise de RSO, que pode ser útil ao nível da tomada de decisão. Antunes e Costa (2012a) analisaram o estado da arte referente à interligação entre as RSO e os SAD e concluíram que existiam pontos de contacto entre ambas as áreas de investigação. Para Power e Phillips-Wren (2012) a *web* 2.0 é a maior inovação tecnológica de apoio à decisão para os gestores.

A enorme quantidade e diversidade de dados produzidos pelas RSO enquadra-os no conceito de *big data* (Davenport, 2014; Khatri et al., 2014; Stieglitz et al., 2018; Zadrozny e Kodali, 2013). Este conceito encontra-se associado aos dados que não podem ser armazenados, processados e analisados, através dos meios tradicionais devido ao seu volume, variedade e velocidade. De acordo com Bahga e Madisetti (2019), Barlow (2013) e Khatri et al. (2014) , o volume descreve a quantidade de dados, a variedade refere-se ao tipo de dados (estruturados e não estruturados produzidos por muitas fontes) e a velocidade é a frequência, taxa ou ritmo, com que são produzidos.

A literatura atual (Barends e Rousseau, 2018; Davis, 2019; Isson, 2018; Kamki, 2016; Pozzi et al., 2017; Rosenfeld e Kraus, 2018; Russell e Klassen, 2019) liga os dados das RSO ao apoio e à tomada de decisão, mas de forma ainda muito teórica e dentro das áreas da ciência dos dados, *Business Intelligence* (BI)[36], *data-driven model*[37], entre outras. Para Ozyer et al. (2019), a investigação em RSO para apoio à decisão assenta menos nas ciências sociais e suporta-se mais nas abordagens multidisciplinares associadas às ciências da computação.

A gestão contemporânea preocupa-se cada vez mais em decidir bem e à primeira, através de estruturas de decisão, processos e mecanismos que sejam facilmente adaptáveis. Dado que a comunicação e a interação entre os elementos de um grupo são importantes para apoio à decisão, a análise de RSO, bem como os conteúdos que nela são produzidos e partilhados, desempenham um papel importante em todo esse processo.

Todavia, apesar de, para Antunes e Costa (2012b), no início da década passada a utilização das RSO para apoio à decisão ser um assunto relativamente novo, de acordo com a literatura recente (Barends e Rousseau, 2018; Davis, 2019; Isson, 2018; Pozzi et al., 2017; Russell e Klassen, 2019) ainda são poucos os trabalhos de investigação que interligam estas duas áreas. Como referido em Freire et al. (2021), as atividades diárias das empresas podem beneficiar da utilização dos dados das RSO. A análise desses dados traz contribuições para os decisores que atualmente enfrentam grandes desafios para se manterem no mercado e compreenderem os seus clientes e concorrentes.

Antes da existência da *internet*, da democratização e acessibilidade à *web* social, quando as pessoas pretendiam decidir sobre um qualquer tema, recolhiam informação junto de amigos e/ou conhecidos. Faziam-no tanto para decisões pessoais como organizacionais, quer fosse para comprar um carro quer fosse para indagar sobre um candidato para um emprego. Como refere Pang e Lee (2008), "o que os outros pensam" sempre foi uma peça de informação importante para a maioria das pessoas durante o processo de tomada de decisão. As RSO possibilitam, sem inquirir ninguém, recolher informação das opiniões e experiências de um vasto conjunto de utilizadores que nem sequer são nossos conhecidos.

[36] BI: técnicas, SI e TI, práticas, metodologias e *software* que analisam dados organizacionais para auxiliar a empresa a compreender melhor os seus negócios e mercado e com isso decidir melhor (Chen et al., 2012).

[37] *Data-driven model* são modelos de *machine learning* para prever a tomada de decisão humana utilizando dados de um contexto organizacional (Rosenfeld e Kraus, 2018).

Este facto alterou a forma como olhamos para todo o processo de decisão (Freire et al., 2017).

Tollinen et al. (2012) argumentam que, comparativamente com inquéritos e/ou entrevistas, a monitorização da *web* social é uma forma mais objetiva para medir e analisar as intenções sociais. Todavia, de acordo com a literatura (Davis, 2019; Isson, 2018; Savic et al., 2019) ainda existem algumas questões por resolver tais como a gestão dos conteúdos das RSO, a sua heterogeneidade e a extração de dados.

As etapas associadas a todo o processo de análise de dados oriundos das RSO são em grande parte do domínio das ciências da computação. Nesse âmbito, os SI e as TI contribuem para o apoio à decisão através do acesso em tempo útil à informação, das bases de dados para armazenamento e do processamento dos dados para obtenção de resultados fiáveis.

Apesar dos SI e das TI ajudarem a reduzir custos organizacionais (pela otimização de processos e tarefas), acabam por ser pouco eficazes na melhoria e no apoio à tomada de decisão. Isto porque poucas empresas recolhem e analisam os próprios dados. De acordo com Isson (2018), as empresas são as primeiras a admitir que conseguem extrair pouco conhecimento dos dados. Do ponto de vista organizacional, manter uma presença social *online* sem ferramentas adequadas para dela extrair informação e a transformar em conhecimento útil, parece ser um desperdício de recursos (Freire et al., 2021).

Atualmente, a forma de decidir está, por isso, a sofrer uma transformação e a evoluir no sentido da antecipação da resolução dos problemas. Para Filip et al. (2017), a resolução de tais problemas estabelece-se com decisões proativas. As RSO vieram alterar as variáveis organizacionais com implicações e consequências nos processos de tomada de decisão já existentes e um efeito direto nos modelos de decisão. Porém, os modelos de análise de dados sociais beneficiam da incorporação dos modelos tradicionais de decisão, das técnicas e da tecnologia da ciência da computação para pesquisar a *web* social e identificar/explorar possíveis problemas e alternativas, para depois os avaliar.

4.3 Desafios de análise das RSO

A produção diária de novos dados nas RSO associada ao seu volume e à sua diversidade criam uma lacuna entre a sua recolha e a capacidade de os organizar e analisar. Independentemente da tecnologia atual (em termos de *hardware e software*) permitir

armazenamento e acesso de forma eficiente (Isson, 2018; Kantardzic, 2011), essa lacuna acentua-se perante os dados não estruturados. Apesar do volume dos dados ser importante, para Davenport (2014) o problema dos dados está na sua falta de estrutura.

A investigação em RSO tem, por isso, como ponto fraco o processamento semântico dos conteúdos textuais. Devido ao seu formato e à sua complexidade, a literatura (Das, 2016; Davis, 2019; Dunaway, 2018; Isson, 2018; Torres-Moreno, 2014; Xhafa et al., 2015) caracteriza-os como dados não estruturados. Isto deve-se ao facto de os utilizadores terem desenvolvido formas de articular vários recursos linguísticos, transportando-os para o discurso das RSO, tornando a comunicação *online* similar à presencial. Para além disso, as novas tecnologias exigem novas formas de linguagem que expressem os novos conceitos que emergem. Como referem Jokinen e Wilcock (2017), as novas tecnologias afetam a utilização da linguagem e esta adapta-se às mudanças por ela provocadas.

Arranjar forma de estruturar este tipo de dados textuais, remete para a área da linguística (Isson, 2018; Torres-Moreno, 2014), remetendo ainda para o domínio da semântica e, em particular, para a *web* semântica. Todavia, a sua estrutura ainda não responde às necessidades de contextos organizacionais específicos, visto que seria necessário, como referem Antunes et al. (2016), criar, de forma completa e exaustiva, padrões para todos os conceitos possíveis e existentes na *web*.

O processamento de dados textuais apresenta, por isso, desafios e limitações de vários níveis, tais como lidar com a codificação de caracteres até inferir significado de um contexto específico (Ingersoll et al., 2013; Isson, 2018). Apesar das várias limitações, associadas ao processamento, estruturação e recuperação de informação, que não podem ser descartadas, a tarefa de extrair e analisar os conteúdos das mensagens das RSO, enquanto dados textuais, é bastante desafiadora por diversos motivos (Freire et al., 2021). Entre eles salientam-se:

- Processamento dos dados textuais - O processamento dos dados textuais contidos nas mensagens, exige ferramentas para limpar e uniformizar os mesmos, no sentido de apreender os aspetos semânticos para que se consiga ir para além da identificação de palavras-chave (Robinson et al., 2015). Se, por um lado, o estilo de escrita utilizado nas RSO tem vulgarmente um padrão fora do comum por vezes incompreensível para aqueles que não fazem parte do discurso *web* ou daquela cultura ou contexto (Antunes et al., 2014), por outro, os

utilizadores cometem erros, ortográficos e/ou gramaticais, utilizam abreviaturas (qd = quando), símbolos (":-)" = risos, ":(" = triste), "esticam" as palavras (nãoooooooo), imagens, áudio, etc. Assim, as ideias contidas nas mensagens podem ser difusas e as frases longas e cheias de caracteres de pontuação, a enfatizar as situações. Nestas mensagens encontram-se também frases incompletas que terminam com reticências ("…"), muitos espaços em branco e gírias. Essas características condicionam qualquer tentativa de utilizar ferramentas já existentes de PLN (*thesaurus*, dicionários, ontologias, etc.) ou qualquer análise linguística suportada por técnicas das ciências da computação;

- Características semânticas - Os conteúdos semânticos das RSO não partilham de ontologias comuns. Se por um lado há trocas discursivas que não contêm sequer texto, apenas *links* e/ou *tags*, por outro encontram-se escritas em mais do que um idioma (*code-mixing*[38]). A utilização de *tags* para classificar temas é um mecanismo muitas vezes ambíguo, onde os utilizadores atribuem significados estranhos para identificar uma situação ou um objeto. Esse vocabulário não controlado e utilizado em *folksonomies*, segundo Antunes et al. (2014) e Freire et al. (2017), deturpa a gestão e codificação dos dados;
- Dimensão das mensagens - De uma forma geral, os utilizadores das RSO, com o objetivo de acelerarem a comunicação, tendem a reduzir a quantidade de caracteres digitados para expressar uma ideia. Os utilizadores omitem também muita informação, para que o discurso *web* seja um resumo do que pretendem dizer, o que pode dificultar a contextualização dos conceitos (Antunes et al., 2014). Em sentido oposto, dependendo da RSO, em alguns casos as mensagens caracterizam-se por serem muito extensas, com uma linguagem cuidada e formal, com poucas abreviaturas ou outras formas de comunicação tais como *smiles*. No *Facebook*, em especial, são permitidos mais de 60.000 caracteres por *post*. O facto de não existir limitação na quantidade de caracteres inseridos nos *posts* pode levar à produção de grandes volumes de dados, incorporando maior dificuldade na sua interpretação. No caso do *Twitter* as mensagens encontram-se limitadas a 280 caracteres, o que simplifica o seu processamento, mas podendo aumentar mais omissões, acrónimos e abreviaturas;

[38] O termo *code-mixing* refere-se à utilização de vocabulário e gramática característicos de idiomas distintos e que aparecem na mesma frase (Bodomo, 2010).

- Características dos dados - Os dados provenientes das RSO são muito mais do que um simples conjunto de ligações e textos. São redes complexas, interativas e dinâmicas, constituídas por múltiplas ligações que correspondem não só aos seus utilizadores e "amigos" (seguidores), mas também às ligações entre mensagens, fotos, etc. (Herring, 2010; Robinson et al., 2015; Velde et al., 2015). Para além disso, os utilizadores têm a possibilidade de comentar mensagens de forma sucessiva e de as partilhar. Esse facto levanta questões de análise, representação e visualização gráfica. As características dos dados recolhidos são por isso de três tipos: estruturados, semiestruturados e não estruturados (os conceitos contidos nas mensagens).

Este novo tipo de dados, característico das RSO, afeta não só a sua interpretação e análise, mas também a sua recolha, armazenamento e processamento. Por essa razão, tão ou mais importante que os motivos acima referidos, são a extração e armazenamento dos dados das RSO. Em particular, a dimensão das mensagens tem influência direta na recolha e armazenamento desses dados.

A extração de grandes volumes de dados pode ser uma tarefa árdua e muito dependente do desempenho das máquinas (computador), do *software* utilizado e das respetivas configurações (32 *bits* vs 64 *bits*; *RAM*; configuração da capacidade do *Gephi*). Segundo McColl et al. (2014), é possível armazenar grandes grafos em disco, mas os algoritmos para os processar sobrecarregam os sistemas computacionais. Este facto, ligado à capacidade de processamento dos dados, pode colocar em causa a análise dos conteúdos do discurso *web* em tempo útil, embora permitindo uma análise *ex post facto*.

O armazenamento dos dados pode também ser uma tarefa árdua, pois envolve técnicas e conhecimento do domínio das bases de dados para auxiliar o seu processamento e representação, mesmo para os gestores que possuem competências nesta área. Recentemente, verificou-se um aumento do interesse pelos dados sociais por parte das ciências da computação ligadas ao armazenamento de dados. O que justifica esse aumento é a necessidade de lidar com grandes volumes de dados (Kroenke et al., 2017) e a procura de técnicas e metodologias para extrair conhecimento do comportamento social, que existe nos dados da *web* social. Alguns exemplos da literatura nessas áreas são Kroenke et al. (2017), Meier e Kaufmann (2019), Perkins et al. (2018) e Robinson et al. (2015), entre outros, que descrevem estruturas de armazenamento de dados para as RSO.

Segundo Kamki (2016), a gestão do volume de dados é ainda um desafio para as empresas. Mesmo recorrendo a *software* de utilização comum, é necessário inferir dados (através de ferramentas construídas para o efeito), recorrendo ao conhecimento e técnicas das bases de dados (relacionais e *graph database*). A análise de dados em tempo real é, por isso, um desafio que pode ser abordado através de janelas temporais utilizando *snapshots*. Esta abordagem permite reduzir o volume de dados (quantidade de mensagens) recolhidos em cada momento, agregando-os *a posteriori*. Permite ainda uma análise constante e sistemática do discurso *web* garantindo que os dados e a informação não ficam desatualizados e possam servir para sustentar a tomada de decisão sem a distorcer.

4.4 SNA na tomada de decisão

O modelo de decisão proposto por Simon (1977) resume o processo de decisão em quatro fases. Em cada fase, para efeitos de apoio à decisão organizacional, o modelo pode incluir a utilização de ferramentas e métodos quer da perspetiva organizacional, quer da perspetiva tecnológica. A combinação de métodos qualitativos e quantitativos da SNA pode evidenciar e caracterizar, não só a estrutura das RSO, mas também as interações existentes entre indivíduos e a sua posição estratégica dentro da rede. O valor dos dados das RSO pode ser revelado com a SNA, que oferece informação e oportunidades para auxiliar os processos de apoio à decisão.

As RSO permitem a extração direta das opiniões e experiências de um conjunto alargado de pessoas, sem necessidade de inquéritos formais (Freire et al., 2017). A literatura referente à aplicação da SNA em RSO, para apoio à decisão, encontra-se fragmentada e não oferece uma visão completa dos elementos que permitem às organizações apoiar as várias fases dos processos de decisão. Muitos estudos na área das RSO lidam apenas com a análise de dados, que a literatura descreve como sendo o processo de analisar e inspecionar dados para apoio à decisão. Esses estudos não incorporam as restantes etapas do modelo proposto (extração, processamento e análise dos dados) necessárias à criação de *outputs* de dados com informação útil. Todavia, a análise de dados é apenas uma parte do processo de análise de RSO. A incorporação e interligação dessas etapas, nas fases do processo de decisão, permitem um acompanhamento dinâmico de um acontecimento e, eventualmente, alterar o curso das decisões de acordo com os dados obtidos. As fases do modelo de decisão proposto por Simon (1977) relacionam-se com a SNA e podem mudar a forma como encaramos o processo de decisão.

A extração de dados das RSO pode relacionar-se com a fase de inteligência que envolve a procura de informação e inclui atividades que visam identificar situações, problemas e/ou oportunidades. Nesta fase, os dados sociais são úteis enquanto fonte de informação que apoia a resolução de um problema. Pode, por isso, considerar-se que a fase de inteligência é o processo de extração e armazenamento de dados de um contexto específico para o reconhecimento e classificação de um problema ou decisão. Nesta fase, os dados das RSO são recolhidos e, quando agregados, podem ser úteis na identificação de informação implícita. As técnicas de visualização da SNA permitem filtrar os dados de uma rede global ou de diferentes sub-redes, convertendo-os em algo mais compacto e compreensível.

Na fase de projeto, as métricas de centralidade da SNA e a análise de várias redes permitem uma visão mais estruturada dos dados do discurso *web*, identificação de informação relevante que pode servir como base para a obtenção de alternativas ou cursos de ação, podendo levar a uma melhor compreensão do problema. As métricas da SNA são úteis, nesta fase, pois permitem não só identificar um problema, mas também investigar alternativas para a sua resolução. Isso significa que as métricas da SNA podem indicar alternativas para possíveis processos de decisão, isolando a informação mais relevante. A incorporação da SNA nesta fase representa a capacidade de encontrar valor a partir dos dados recolhidos, com o objetivo de obter alternativas relevantes para resolução de problemas organizacionais.

Escolher envolve quase sempre incerteza, ou porque remete o decisor para ações que serão executadas no futuro, ou porque não se fundamenta em conhecimento ou princípios absolutos, mas sim em suposições, perceções e/ou opiniões fundamentadas em valores pessoais ou crenças. Assim, na fase de escolha, as técnicas matemáticas e de visualização gráfica da SNA utilizam-se para identificar e escolher possíveis alternativas "escondidas" nos dados das RSO. A análise e visualização permitem, por um lado, evidenciar conceitos relevantes do discurso *web* (rede semântica) e, por outro, as métricas associadas às interações entre utilizadores podem revelar características importantes e auxiliar uma tomada de decisão informada. Nesta fase processa-se a escolha de alternativas, utilizando a análise de dados feita na fase anterior, o que permite aos decisores obter uma visão global dos problemas em análise. Em seguida, a partir das análises efetuadas com as técnicas da SNA, opta-se por uma ou várias das alternativas criadas anteriormente para resolver o problema de decisão. A análise e visualização das RSO pode ser usada na fase de

escolha, porque é o momento crítico da tomada de decisão e é o ponto em que se decide por um curso de ação. Por exemplo, a análise e visualização de conceitos utilizados no discurso *web* podem ser utilizadas como conhecimento para decisões de acompanhamento, pois permitem acesso eficiente às informações principais. Também nesta fase, o nível de interpretação das ligações dos utilizadores pode ser o número de relacionamentos que um determinado utilizador tem com outras pessoas ou o número de *links* que ele possui para uma mensagem. Avaliar esses factos pode ser extremamente importante, pois pode ajudar a tomar decisões mais esclarecidas, em áreas de negócios como, por exemplo, a comunicação, as relações públicas, o *marketing* digital, a gestão de conteúdos *web*, etc. Nesta fase, são também definidos critérios e regras que os decisores utilizarão para avaliar as várias alternativas que melhor se ajustam ao problema em análise.

Por último a fase de implementação e avaliação do processo de decisão é onde a alternativa escolhida é executada. Envolve pôr em funcionamento uma solução encontrada nas etapas anteriores. Após identificar alternativas e a sua aplicação, com critérios e regras, os resultados da SNA auxiliam os decisores na monitorização do comportamento da(s) rede(s), comparando métricas anteriores e subsequentes. Portanto, a SNA pode ajudar os decisores a ver e perceber a melhor forma de implementar opções para resolver um problema específico. Por exemplo, no discurso *web*, os resultados obtidos da SNA fornecem informação referente às palavras-chave utilizadas durante as trocas discursivas e podem utilizar-se para analisar várias perspectivas sobre um problema e apoiar a tomada de decisão. Isso significa que, nesta fase, o decisor pode utilizar aqueles resultados, obtidos da SNA na fase de escolha, para concluir o processo de tomada de decisão com a seleção de uma ou mais alternativas, que podem ser implementadas pela organização.

Em resumo, tudo o que é produzido na *web* social é possível de ser capturado, armazenado e analisado para apoiar a tomada de decisão e pode estar disponível para ser utilizado a qualquer momento. A SNA, associada ao processo de decisão, desempenha um papel importante no contexto organizacional, com ênfase na criação de alternativas para o processo de apoio à tomada de decisão. Ao vivermos num mundo global permanentemente ligado e em que as preferências do consumidor mudam a todo o momento, as organizações precisam de monitorizar os canais de comunicação, como as RSO, para perceber os perfis e/ou comportamentos dos consumidores e as métricas de centralidade descritas no capítulo 3, podem ser consideradas para identificar as propriedades mais significativas dos nós das entidades do discurso *web*. Estas métricas indicam o prestígio, poder, influência e importância

não só dos utilizadores, mas também das suas trocas discursivas e são indicadores importante para as quatro fases do processo de decisão.

4.5 Estudos em RSO

4.5.1 Evolução

É reconhecido que a investigação que utiliza dados da *web* social para apoio à decisão, ainda está numa fase inicial e pouco desenvolvida (Alhajj e Rokne, 2018; Davenport, 2014; Stieglitz et al., 2018), em particular no que se refere à análise semântica dos conteúdos das mensagens. A investigação em RSO envolve a análise de dados de plataformas da *web* social, tais como *Facebook*, *Twitter*, *Instagram*, *WhatsApp*, entre outras, que guardam os conteúdos produzidos pelos seus utilizadores. Se considerarmos que o *Facebook* foi lançado em 2004, o *Twitter* em 2006, o *WhatsApp* em 2009 e que o *Instagram* foi fundado em 2010, podemos afirmar que esta área de investigação está no seu estágio inicial. Os dados produzidos pelos próprios utilizadores incluem os seus perfis, publicações e comentários, padrões de interação, bem como as múltiplas redes que daí emergem e por isso têm atraído muitos investigadores e profissionais.

Nestes últimos 15 anos, os trabalhos de investigação que utilizaram os dados das RSO acima referidas são essencialmente teóricos e têm o seu foco mais nas características estruturais das redes do que nos aspetos semânticos dos seus conteúdos (Freire et al., 2017). A maioria dos estudos limita-se à descrição das propriedades da rede de utilizadores e aborda superficialmente a análise semântica das mensagens, pois para o efeito é necessário entrar no domínio da linguística, do discurso, do PLN, domínios por si só complexos (Freire et al., 2021). Na literatura encontram-se estudos mais complexos que aprofundam a análise semântica dos conteúdos do discurso *web*, mas não interligam com a rede de utilizadores.

Inicialmente, entre 2005 e 2010, devido à dificuldade em adquirir e estruturar dados de interação social, alguns investigadores optaram por criar plataformas sociais (fechadas) de argumentação. Assim, conseguiam à partida estruturar sequencialmente as trocas discursivas do discurso *web* e trabalhar com dados previamente estruturados para apoio à decisão. Várias plataformas foram propostas para analisar, em conjunto, o relacionamento social e os conteúdos produzidos nas RSO para apoio ao processo de tomada de decisão. Schneider et al. (2010) realizaram um levantamento dessas plataformas, 37. Shum (2008) analisou 10 plataformas de argumentação, quer de um único utilizador quer colaborativas, e

apresentou uma, o *Cohere*. O *Cohere* desenvolvido para apoio à decisão por Shum et al. (2011) é uma plataforma *web* 2.0 de análise e visualização de argumentos que permite a decisão colaborativa *online*. Por seu lado, Scheuer et al. (2010) analisaram 50 sistemas de argumentação e Rahwan et al. (2007) analisaram 12. Para além do *Cohere* já referido, outros exemplos dessas plataformas são o *Parmenides* (Atkinson et al., 2006), o *Smartocracy* (Rodriguez et al., 2007), o *Impact* (Balk, 2008), o *DebateGraph* (Price et al., 2011) entre outras.

Entre 2010 e 2015, com o acesso a ferramentas de fácil utilização, tais como o *Netvizz* e o *NodeXL*, desenvolvidas no meio académico para extração, visualização e análise de dados, verificou-se um aumento considerável da investigação em RSO. Nesse período, como referiu Russell (2014), as possibilidades de análise de uma página do *Facebook* passaram a ser praticamente ilimitadas e, depois da análise das interações entre utilizadores, o passo seguinte seria analisar os dados da linguagem humana. Todavia, os dados textuais das mensagens ainda são de difícil análise visto que não se encontram num formato padronizado e requerem um pré-processamento para que possam ser utilizados. A literatura (Das, 2016; Davis, 2019; Dunaway, 2018; Isson, 2018; Torres-Moreno, 2014; Xhafa et al., 2015) caracteriza-os como dados não estruturados.

Neste período, a capacidade cada vez mais facilitada de recolher e analisar os conteúdos das mensagens abriu novas linhas de investigação. Passou a ser possível estudar a interação social entre utilizadores, bem como ir um pouco mais além e estudar as características textuais dos conteúdos de uma mensagem, ou seja, a sua semântica. Em particular, o grande volume de dados produzido nas RSO é de uma riqueza tal, quer para a comunidade académica, quer para as empresas e organizações, que permite compreender o comportamento *online* e monitorizar tendências sociais para o apoio à decisão. Todavia, apesar do aumento de trabalhos de investigação nesta área, fruto do acesso facilitado à extração de dados, verifica-se que muitos deles são estudos de caso isolados, em que foi recolhido um conjunto de dados de um período temporal específico, de um tema específico e analisado com recurso à SNA.

No seu estágio inicial, a investigação estava voltada maioritariamente para a aplicabilidade e a análise das RSO numa área específica e por acréscimo às suas questões de investigação, e não para as técnicas e métodos necessários para abordar essas questões. Atualmente muitos estudos ainda lidam apenas com a análise de dados, que a literatura (Bassan e Sarkar, 2014; Nisbet et al., 2018; Roiger, 2017) descreve como sendo o processo

de extração de informação relevante para apoiar a tomada de decisão. Contudo, o processo de análise das RSO é composto por várias fases, das quais a análise de dados é apenas uma delas.

Recentemente, a investigação em RSO passou a fundamentar-se menos nas ciências sociais e mais nas ciências da computação, para poder lidar com métodos interdisciplinares ligados à análise de dados (Moreira et al., 2019; Stieglitz et al., 2018). Os motivos que justificam essa mudança são vários. Se, por um lado, o grande volume de dados produzidos *online* impulsionou a investigação para além de seu limite inicial, a investigação das interações sociais, passando a incluir também a análise do conteúdo semântico, por outro lado, as consequências das novas realidades económicas e sociais e o acesso livre a dados sociais de acontecimentos do mundo real, deram início a uma nova era de investigação e desenvolvimentos na análise de RSO (Campbell et al., 2013). A *web* social e as RSO representam uma importante fonte de informação, quer para as empresas, quer para os académicos pois, como refere a literatura (Barends e Rousseau, 2018; Davis, 2019; Isson, 2018; Osei-Bryson e Rayward-Smith, 2009; Power et al., 2011; Pozzi et al., 2017; Russell e Klassen, 2019), simplificam e estimulam a criação de grandes volumes de dados.

4.5.2 Análise de conteúdo/semântica

Os estudos que incorporam a SNA, bem como a análise semântica das mensagens, são escassos e principalmente teóricos. Nestes, a análise do discurso *web* produzido nas RSO é abordada do ponto de vista da CMC (Herring, 2010, 2013), da análise do discurso (Herring, 2010; Moser et al., 2013), da análise de conteúdos *web* (Kok e Rogers, 2016) ou do PLN (Agogo e Hess, 2018; Bharti et al., 2017). Para analisar o texto produzido no discurso *web*, esses estudos focam-se não só nos aspetos discursivos, retóricos, semânticos e pragmáticos, mas também nos métodos formais de análise de texto do PLN. Nesse sentido, verifica-se a existência de uma transferência de conceitos e conhecimento entre áreas, que ajuda na compreensão dos fenómenos mais recentes da comunicação *online* e o consequente discurso argumentativo associado.

Outros estudos que efetuam análises semânticas das mensagens das RSO, normalmente adotam uma abordagem manual para tratamento e codificação do texto. Erétéo et al. (2011) conjugaram a estrutura da rede de utilizadores e o conteúdo por eles produzido, através de *tags*. Os autores adicionaram manualmente *tags* às mensagens, para estabelecer e identificar as relações entre elas, bem como entre as mensagens e os utilizadores. Para a

análise semântica, Moser et al. (2011) utilizou os *posts* dos três utilizadores mais centrais da sua rede, do *Facebook*, e analisou manualmente e individualmente as mensagens de cada um. Para o efeito utilizou a análise de *co-word*[39]. No trabalho de Himelboim et al. (2013), que também interligou a SNA e os conteúdos do *Twitter*, os 212 *tweets* foram codificados manualmente. Velde et al. (2015) combinaram as características das mensagens e dos utilizadores através de um modelo linear multidimensional, em que os *tweets* foram codificados manualmente por um investigador principal, utilizando uma lista predefinida. Num estudo de *marketing*, Kefi et al. (2016) identificou o conteúdo temático de diferentes comunidades da RSO, analisando manualmente a informação contida em cada publicação. No trabalho de Réda et al. (2017), os 153 *tweets* analisados para construção de um sistema de recomendação, que identifica informação relevante para especialistas, foram codificados manualmente. Os autores extraíram e classificaram, também manualmente, as palavras-chave de um terço dos *tweets* que codificaram (50 de 153).

Os sistemas de recomendação são também uma área de investigação importante. Diversos autores (Gou et al., 2011; He e Chu, 2010; Ma e Che, 2016; Samanthula e Jiang, 2014) analisaram as interações entre utilizadores e as suas mensagens para criarem algoritmos de recomendação. Apesar de todos utilizarem a abordagem semântica para análise de conteúdo, restringem-se à identificação de palavras-chave ou tópicos.

Por questões de dificuldades na análise e estruturação dos dados do discurso *web*, alguns investigadores restringem os textos que analisam a um domínio de conhecimento específico, visto que já estará à partida padronizado em bases de dados de texto. De acordo Zhai e Massung (2016), como os dados de texto são a forma mais natural de codificar o conhecimento humano, pois, por exemplo, o conhecimento científico existe quase exclusivamente na literatura científica, enquanto que o conhecimento técnico em manuais técnicos. Por exemplo, Saint-Dizier (2012) restringiu o seu estudo a textos técnicos para analisar as opiniões dos clientes e identificar os motivos de satisfação ou insatisfação de um determinado produto ou marca. De acordo com o autor, precisamente porque são muito mais fáceis de processar comparativamente aos textos informais, característicos da *web* social. Para além disso, como refere o autor, a importância dada às características semânticas e

[39] Técnica da análise de conteúdo que evidencia palavras-chave em dados textuais (Leydesdorff e Hellsten, 2006).

pragmáticas fica dependente do tipo de problema investigado, do género textual utilizado e do binómio orador/público-alvo.

O trabalho de Beck-Fernandez et al. (2017), para além de apresentar uma base teórica, propõe um sistema de processamento e extração de texto e a sua conversão em *memes*[40]. Os autores resumem o conteúdo textual e não estruturado de mensagens escritas em linguagem informal, construindo redes semânticas e identificando os temas principais e mais representativos do texto como unidades de conhecimento transmissível (*i.e. memes*). O trabalho de Biswas et al. (2018) também propõe um método sem supervisão (*unsupervised*[41]) de extração de palavras-chave dos conteúdos do *Twitter,* combinando a visualização gráfica e métricas da SNA. Os autores elegem as palavras-chave combinando a frequência, centralidade, posição e força dos vizinhos de um nó, para calcular a importância do mesmo. Em sentido oposto, Duari e Bhatnagar (2020) para construção de redes semânticas e identificação de palavras-chave utilizam bases de dados com textos (*abstracts* e artigos) de domínios científicos e de notícias publicadas *online*. Os autores utilizaram listas predefinidas e já normalizadas de palavras-chave, ou seja, utilizam um método com supervisão (*supervised*).

Recentemente verificou-se, por parte das ciências da computação ligadas ao armazenamento de dados e *data-mining,* um aumento do interesse em dados sociais. A principal motivação é a procura de técnicas e metodologias para explorar e extrair conhecimento do comportamento social, que existe nos grandes volumes de dados da *web* social. Alguns exemplos são a quantidade de literatura nessas áreas (Ravindran e Garg, 2015; Russell e Klassen, 2019; Szabo et al., 2018) que evidencia os esforços de investigação em RSO. As técnicas do *data-mining* têm-se mostrado úteis para tratar grandes volumes de dados que não poderiam ser manipulados pelos métodos tradicionais. Muitos dos estudos realizaram análises semânticas das mensagens das RSO e utilizaram a SNA no processo de *data-mining* e *text-mining* (Bhanap e Kawthekar, 2015; Fernando et al., 2015; Joseph et al., 2016; Pippal et al., 2014; Sheshasaayee e Jayanthi, 2015).

[40] *Meme,* conceito utilizado para descrever ideias, imagens, vídeos, etc. partilhados via *internet* (Lima, 2011).

[41] A extração de palavras-chave divide-se em duas categorias: *supervised* e *unsupervised* (Isson, 2018). Na primeira, utiliza-se um algoritmo para extrair palavras-chave de acordo com variáveis de entrada "x" (conceitos previamente padronizados) e o resultado é uma variável de saída "y". Na segunda, *unsupervised*, não há uma comparação e correspondência dos dados de entrada "x" e variáveis de saída. O objetivo é modelar uma estrutura, por exemplo *cleaning database,* com conceitos existentes nos dados para melhor os entender.

Outros estudos, para a análise das redes semânticas, utilizaram *word clouds*[42], em substituição das métricas da SNA, para visualizar o nível de importância das palavras e identificar palavras-chave no texto (Chatterjee e Trumbo, 2018; Ghim et al., 2018; Kleminski e Kazienko, 2018; Tripathy et al., 2017; Troisi et al., 2018).

Alguns trabalhos investigaram em exclusivo a problemática da estruturação de dados semânticos. Por exemplo, Tiroshi et al. (2017), investigaram só a problemática associada à transformação dos dados não estruturados dos conteúdos da *web* social. O modelo que os autores propuseram estrutura dados semânticos (*posts*) em formato tabular a partir da perspetiva da teoria dos grafos e utiliza as métricas da SNA para a extração de novos dados.

4.5.3 Utilização de grafos

Nas últimas décadas, as comunidades científicas ligadas à SNA, linguística e computação têm introduzido novos conceitos, algoritmos e aplicabilidades dos grafos ao mundo real e em particular para o apoio à decisão no contexto das RSO. Alguns autores (Gordon, 2011; Harrison, 2015; Kemper, 2015; Laudon e Laudon, 2011; Perkins et al., 2018; Robinson et al., 2015) tentaram dar resposta à problemática do armazenamento dos dados sociais e à sua representação gráfica, utilizando quer técnicas das bases de dados quer da teoria dos grafos. Nestas áreas destaca-se um novo tipo de bases de dados gráficas, as g*raph databases,* que lidam com dados altamente interligados entre si.

Nesse sentido, alguns trabalhos investigam a problemática do armazenamento dos dados sociais em formato grafo, quer seja para armazenamento de dados estruturados, interações entre utilizadores, quer seja para análise de redes semânticas com dados não estruturados. Campbell et al. (2013) abordam os problemas da construção de redes de dados não estruturados, analisando a estrutura da comunidade da rede e inferindo informação a partir daí. Este trabalho teórico descreve como a aplicação de ferramentas e metodologias tais como bases de dados, análise gráfica e visualização, *information extraction,* são importantes na análise de dados não estruturados. De acordo com os autores, a representação do conhecimento através de *graph databases* segue uma abordagem equivalente à utilizada nos modelos de bases de dados relacionais. Para os autores, as *graph databases* são uma

[42] *Word clouds* ou *tag cloud*: O tamanho da visualização de um conceito é proporcional à sua popularidade ou importância e organizam-se aleatoriamente ou por ordem alfabética dentro de uma "nuvem" (Croft et al., 2015).

ferramenta valiosa para resolver os desafios dos dados sociais (isto é, as suas características, volume, etc.).

O trabalho teórico realizado por McColl et al. (2014) apresenta um estudo qualitativo e compara o desempenho de 12 *graph databases*. Para os autores, muitas das abordagens de *graph databases* desenvolveram-se tendo por base a longa história e investigação feita no domínio das bases de dados relacionais. Para medir o desempenho das 12 bases de dados, os autores utilizaram quatro conjuntos de dados com volumes diferentes (minúsculo, pequeno, médio e grande) aos quais aplicaram quatro algoritmos gráficos.

4.5.4 Apoio à decisão

A literatura (Barends e Rousseau, 2018; Davis, 2019; Isson, 2018; Osei-Bryson e Rayward-Smith, 2009; Power et al., 2011; Pozzi et al., 2017; Russell e Klassen, 2019) é unânime ao salientar que as RSO são importantes para o apoio à decisão, mas poucos são os trabalhos de investigação que interligam estas duas áreas. Neste âmbito, Sadovykh e Sundaram (2016, 2017) e Sadovykh et al. (2015a, 2015b), analisaram o discurso *web* produzido pelos utilizadores de RSO, utilizando a metodologia da netnografia em vez de recolher de forma automática os dados. Os autores utilizaram essa metodologia para explorar o potencial das RSO como ferramenta de suporte ao processo de tomada de decisão. Para o efeito, relacionaram as fases do modelo de decisão de Simon (1977) com a informação observada em RSO, das áreas da saúde e das finanças. A netnografia caracteriza-se por ser o investigador a recolher os dados das interações e trocas discursivas de membros de um grupo da *web* social, através da sua observação e posterior registo digital. Os autores selecionaram 29 sites com conversas, que relacionaram com o processo e fases de tomada de decisão. Consideraram um total de 51 conversas com uma média de 878 palavras por mensagem e uma média de 17 comentários associados a cada uma.

De forma a identificar diferenças entre o modelo proposto neste trabalho de investigação e modelos existentes na literatura para análise de RSO, foram analisados 19 trabalhos publicados na última década. Esses trabalhos mostram a crescente atenção que tem sido dada à análise dos dados das RSO para apoio à tomada de decisão organizacional. A análise incluiu os artigos selecionados e apresentados na Tabela 3. Alguns deles limitam-se a apresentar modelos teóricos (Aladwani, 2014; Alkhyeli e Mansour, 2015; Caroleo et al., 2015; Fernando et al., 2015; Madan e Chopra, 2015; Sarker et al., 2015; Walha et al., 2017; Zielinski et al., 2013) e outros são sobretudo abordagens exploratórias e padronizadas. Esses

trabalhos foram analisados de acordo com cinco características: RSO, *software*/plataforma, entidade(s) de análise, etapas do processo e técnicas SNA.

Tabela 3 - Caracterização dos modelos de análise de RSO encontradas na literatura

(1) Autor	(2) RSO	(3) *Software*/Plataformas	(4) Entidade de análise	(5) Etapas	(6) SNA
Gjoka et al. (2011)	*Facebook*	*HTML scraping*	*User*	*data-extraction data-processing data-analysis*	Não
Zielinski et al. (2013)	*Twitter*	*Twitter Streaming API, MySQL, Apache QPID, Weka, NLTK toolkit*	*User, Post*	*data-extraction data-processing data-analysis*	Sim
Oussalah et al. (2013)	*Twitter*	*Twitter Streaming API, MySQL, Apache Lucene, WordNet, PostGIS, Django, Python*	*User, Post, Semantic content*	*data-extraction data-processing data-analysis*	Não
Aladwani (2014)	*Facebook*	-	*User, Post*	*data-extraction data-processing data-analysis*	Não
Vosecky et al. (2014)	*Twitter*	*URL link, Twitter REST API*	*User, Post, Semantic content*	*data-extraction data-processing data-analysis*	Não
Alkhyeli et al. (2015)	*Twitter, Inquérito*	-	-	-	Não
Banica et al. (2015)	*Twitter*	*Apache: Flume, Sqoop and Hadoop, NoSQL, Gephi, NodeXL*	*User*	*data-extraction data-processing data-analysis*	Sim
Caroleo et al. (2015)	*Twitter, Facebook*	*Cassandra, NoSQL databases, MongoDB*	-	*data-extraction data-processing data-analysis*	Não
Fernando et al. (2015)	-	*Social network APIs*	*Semantic content*	*data-extraction data-processing data-analysis*	Não
Lai et al. (2015)	*Webpages*	*WordSmith, Leximancer, SPSS*	*Semantic content*	*data-extraction data-processing data-analysis*	Não
Madan et al. (2015)	*Facebook*	*JVM (Java)*	*User*	*data-extraction*	Não
Sarker et al. (2015)	Revisão de literatura	-	-	*data-extraction data-processing data-analysis*	Não
Sathick et al. (2015)	BSAU	*R-tool, SQL*	*User, Semantic content*	*data-extraction data-processing data-analysis*	Não
Ghafoor et al. (2016)	Inquérito	-	*User*	*data-extraction*	Sim
Ma et al. (2016)	Questionários, http://www.yelp.com	*JVM (Java), Xpath (XML)*	*User, Semantic content*	*data-extraction data-processing data-analysis*	Não
Vicario et al. (2017)	*Facebook*	*Facebook Graph API, IBM WatsonTM AlchemyLanguage service API*	*User, Pages*	*data-extraction data-processing data-analysis*	Não
Walha et al. (2017)	*Twitter*	*Thomson Reuters Open Calais, Dbpedia*	*User, Post, Semantic content*	*data-extraction data-processing data-analysis*	Não
Appel et al. (2018)	*Claims database*	-	*User (physicians, patients), health providers*	*data-processing data-analysis*	Sim
Ruas et al. (2019)	*Facebook*	*ELKI, SPSS, Gephi, NodeXL*	*User, Post*	*data-extraction data-processing data-analysis*	Não

A coluna dois, especifica as RSO ou o tipo de plataforma da *web* social a partir da qual os dados foram extraídos, isto é, a fonte de dados que o modelo utilizou. Quando

aplicável, identifica se os dados foram extraídos do *Facebook* do *Twitter* ou de outra RSO. O tipo de ferramentas, *software*(s) e/ou plataforma(s), utilizadas pelos autores para extrair, processar e/ou analisar os dados são identificadas na coluna três. Na coluna quatro identificam-se quais as entidades analisadas e incluídas nos modelos. Essas entidades (o utilizador, a mensagem e o conceito) identificam que tipo de dados foram utilizados, ou seja, se os modelos utilizaram dados estruturados, semiestruturados ou não estruturados. Para além disso, a identificação e caracterização das entidades permitiu avaliar se os modelos só analisaram a estrutura da rede de utilizadores ou se incluíram também os conteúdos e/ou os conceitos. Por outras palavras, apesar de nem todos os trabalhos referirem que tipo de entidade investigaram, identificou-se se analisaram ou não as interações entre utilizadores, bem como o conteúdo semântico das mensagens. A análise dos dados das RSO tem três etapas fundamentais que são: *data-extraction*, *data-processing* e *data-analysis*. Os modelos analisados também foram caracterizados, na coluna cinco, de acordo com as etapas necessárias para executar todo o processo de análise de dados das RSO. Por fim, na última coluna, identificou-se se a SNA foi ou não aplicada, isto é, se os modelos utilizaram ou não as métricas da SNA.

Salientam-se a seguir algumas características dos trabalhos indicados na Tabela 3:

- O modelo apresentado por Gjoka et al. (2011) teve como objetivo caracterizar a RSO de acordo com propriedades estruturais e topológicas e obter uma amostra uniforme de utilizadores, caracterizando-os de acordo com as suas propriedades individuais;
- O artigo de Zielinski et al. (2013) descreve um modelo que monitorizou e classificou *tweets* relevantes em situações de crise, constituído por quatro módulos: extração de conteúdos, análise de credibilidade dos utilizadores; análise de georreferenciação dos *tweets*; e classificação multilingue dos *tweets*;
- O trabalho de Oussalah et al. (2013) descreve uma arquitetura para a extração de mensagens do *Twitter* de uma região predefinida, que investigava a arquitetura de *software* da plataforma *Twitter* e apresentou uma nova arquitetura dedicada à análise semântica e de georreferenciação dos dados do *Twitter*;
- O artigo de Aladwani (2014) descreve um modelo teórico para gestão de conteúdos de RSO, composto por seis módulos que descrevem algumas das tarefas associadas à extração e processamento dos dados sociais;

- O modelo proposto por Vosecky et al. (2014) só faz a modelação do conteúdo semântico, identificando palavras-chave através de *hashtags* e *links* de documentos *web*;
- Banica et al. (2015) apresentam as vantagens de explorar os requisitos dos clientes das RSO para estratégias de *marketing*, pelo que os autores utilizaram o *software Gephi* e *NodeXL* e apresentaram uma comparação dos recursos de cada um;
- Fernando et al. (2015) propõem um modelo para a análise semântica das mensagens das RSO, utilizando a SNA no processo de *data-mining* e *text-mining*;
- Na área do comércio eletrónico, Lai e To (2015) utilizam a análise de conteúdos para desenvolver uma ferramenta de gestão de informação e análise dos conteúdos de várias páginas *web*. Esse trabalho propôs uma metodologia que convertia ficheiros de texto, extraídos da *web* social, em conceitos chave, de modo a que fossem facilmente interpretáveis e visíveis com um mapa conceptual. As páginas *web* analisadas foram identificadas manualmente utilizando as palavras-chave Macau, viagens e *blogs* na barra de pesquisa do *Google*. Após identificação das páginas *web*, extraíram os conteúdos para um ficheiro e utilizaram um *software* lexical (o *WordSmith*) que identificou palavras-chave, comparando os dados que recolheram com os dados de um arquivo predefinido de padrões lexicais. Os autores utilizaram também o *software* de mapeamento lexical *Leximancer* que, através de coocorrências de palavras-chave, identifica temas em bases de dados de texto;
- Sathick e Venkat (2015) apresentam um modelo para implementar um sistema de recomendação *online* para ajudar alunos a escolher o curso universitário pretendido. Os autores estruturaram os dados recolhidos, recorrendo a métodos de indexação utilizando *tags*. Os atributos que utilizaram para indexação dessas *tags* foram os *ID* dos documentos *web*, a frequência dos conceitos e a sua posição nos documentos. Nesse trabalho, os dados foram estruturados e transformados com recurso ao PLN, a técnica *unsupervised learning* de acordo com as semelhanças dos conteúdos dos documentos, ou seja, *data clustering*.

- O trabalho de Ma e Che (2016) enquadra-se na área do *marketing* e analisa as interações entre utilizadores e os seus *posts*, com o objetivo de criar algoritmos de recomendação sobre restaurantes. A análise de conteúdo que os autores realizaram identificou apenas palavras-chave ou tópicos predefinidos nos *posts*, com base nas preferências dos amigos dos utilizadores;
- Ghafoor e Niazi (2016) utilizam dados recolhidos por questionário e propuseram um modelo e os requisitos necessários para projetar novas RSO;
- Vicario et al. (2017) desenvolveram um modelo que combinou a análise de sentimentos e a extração automática de tópicos;
- Appel et al. (2018) apresentam um modelo que utiliza a SNA para modelar e analisar dados de reclamações de seguros de saúde;
- O modelo proposto em Ruas et al. (2019) utiliza a SNA para segmentar o comportamento dos utilizadores em *clusters*. Os autores, com base nos resultados que obtiveram, identificaram e classificaram os utilizadores da RSO de acordo com três perfis: espectador, participante e produtor de conteúdo.

4.6 Resumo

Neste capítulo abordaram-se processos e modelos de apoio à decisão. Na literatura encontram-se vários modelos e teorias referentes ao processo de decisão. Contudo, apesar dos inúmeros conceitos existentes, observou-se que o modelo de Simon (1977) é o mais reconhecido e utilizado. Verificou-se ainda que os modelos clássicos de tomada de decisão estão associados a processos sequenciais, que assentam na ideia de que um problema tem várias alternativas e uma solução para o resolver.

Abordou-se o apoio à decisão no contexto das RSO, evidenciando que a comunicação e a interação entre os elementos de um grupo são importantes para apoio à decisão, a SNA na análise das RSO, bem como os conteúdos que nela são produzidos e partilhados, desempenham um papel importante em todo esse processo.

Foram também abordados os desafios da análise de RSO, nomeadamente o surgimento de novos modelos e sistemas capazes de lidar e interpretar as opiniões enquanto objeto para apoio à decisão. O volume de dados e informação que circula na *web* social torna-a por um lado dinâmica, mas, em simultâneo, caótica. O facto de ser composta por

diferentes estruturas e conteúdos sem uma correlação semântica, aumenta a sua complexidade. Para além disso, a falta de organização da informação nela existente dificulta a sua estruturação. A SNA interligada com as técnicas da ciência da computação reforça a utilidade do discurso *web*, em função dos seus utilizadores enquanto produtores de informação útil para apoio à decisão.

Neste capítulo, abordou-se a SNA na tomada de decisão, que permite que o apoio à decisão passa a ser mais preditivo e não tão fechado em conceitos teóricos. Mais do que entender os fluxos de informação existentes nas RSO, torna-se necessário estruturar essa mesma informação, porque o conhecimento dela extraído pode ser relevante para a tomada de decisão em geral ou, em particular, para um problema específico.

Por fim, descreveram-se alguns estudos realizados no âmbito das RSO, SNA e apoio à decisão. Pela análise desses estudos, verificou-se que estamos perante uma área do conhecimento a que as ciências sociais e as ciências da computação não ficam alheias.

Pela análise desses estudos verificou-se que existe um grande volume de investigações com foco nas RSO. No entanto, os estudos que interligam a análise de RSO e o apoio à decisão são poucos e maioritariamente teóricos. Apesar de muitos estudos investigarem as RSO, poucos são os que utilizam os dados aí produzidos ou a SNA como metodologia de análise. Verificou-se também que, apesar da literatura enfatizar a importância dos dados sociais para o apoio à decisão, são pouco os estudos que interligaram as RSO e os modelos de decisão existentes. Para além disso, poucos são os estudos que analisam em simultâneo as três entidades do discurso *web*, utilizador, mensagem e conteúdo textual, para apoio à decisão. Os trabalhos que analisam as três entidades são restritos na forma como o fazem e só utilizam o conhecimento de uma ou outra área de investigação.

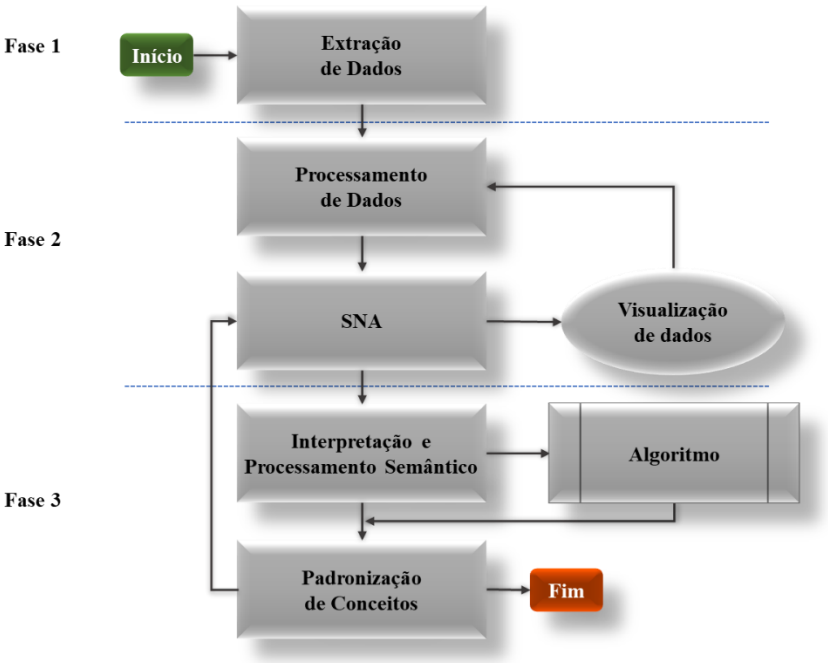
Ao analisar os estudos mencionados, verificou-se também que os desafios associados aos dados sociais não foram abordados na literatura de forma abrangente, o que faz com que estejamos perante uma área multidisciplinar muito fragmentada.

5 Modelo de estruturação e análise de dados das RSO

Neste capítulo descreve-se o modelo desenvolvido e proposto em Freire et al. (2017, 2021) para obtenção de *outputs* de dados estruturados das RSO, para apoio à tomada de decisão. O modelo apresentado neste trabalho de investigação é uma ferramenta de apoio a todo o processo de decisão, mas tendo como principal objetivo a estruturação das interações e do diálogo estabelecido entre utilizadores.

O que diferencia este modelo de outros encontrados na literatura é o facto de articular de forma simples as etapas de extração, armazenamento, processamento e estruturação de dados das RSO, incorporando aspetos importantes tais como a interação humana e a estrutura da rede. Mais precisamente, exploram-se as RSO de duas perspetivas diferentes: a análise das interações entre utilizadores (que difundem a informação) e os diálogos baseados na análise dos textos (que dependem do conteúdo da informação que é transmitida). Como mostrado na Figura 21, o modelo preconiza um início com a recolha de registos das RSO, seguindo-se um processo recorrente e iterativo que produz dados e informação para apoio à decisão.

Figura 21 - Modelo de análise de dados das RSO



O modelo foi desenvolvido e otimizado com recurso a dados de contextos organizacionais reais, extraídos diretamente do *Facebook*. Foi assim possível avaliar e determinar a possibilidade de o modelo final responder às necessidades organizacionais.

A implementação do modelo para recolha das trocas discursivas produzidas na RSO *Facebook*, implicou a utilização do *software NodeXL* e *Netvizz* para extração dos dados e o seu armazenamento em bases de dados criadas em *Excel*. Um dos motivos que levou à escolha destes *softwares* de extração de dados foi o formato simples dos *outputs* que disponibilizavam e que facilmente foram utilizados. O modelo apresentado pode utilizar outras ferramentas de extração de dados, tais como APIs do *Facebook*, para recolha automática e extração de dados em tempo real, mas requer conhecimentos na área da programação.

As bases de dados incorporadas no modelo permitem acesso rápido à informação para que o decisor analise a dinâmica e a estrutura da RSO, as respetivas interações, bem como o conteúdo semântico publicado *online*. Como refere Karacapilidis (2006), uma base de dados é um meio utilizado para organizar informação independentemente da forma como é criada. De acordo com Guerrero (2019) e Morgado (2016), o *Excel* é um dos *softwares* que pode ser utilizado para gerir dados e para Perkins et al. (2018), apesar de menos utilizado, destaca-se quando lida com dados fortemente interligados entre si. Assim, o *Excel* foi utilizado para indexação e manipulação dos dados por três motivos. Em primeiro, porque atualmente muitas organizações o utilizam como ferramenta comum de manipulação de dados. Em segundo lugar, pois, para além de ser uma ferramenta de uso comum, permite gerir e estruturar os dados em formato tabular característico das bases de dados relacionais. Em terceiro lugar, tem a capacidade de facilmente permitir a transição entre análises exploratórias e e/ou versões definitivas de *outputs* de dados. Desta forma, o *Excel* foi utilizado na segunda fase do modelo, como meio para organizar e estruturar informação em base de dados e na terceira para identificar e padronizar as redes de conceitos, ou seja, as redes semânticas.

Na terceira fase foi utilizado também um algoritmo desenvolvido em *Visual Basic for Applications* (VBA)[43] incorporado no *Excel*. Uma das vantagens encontradas na

[43] O VBA é uma linguagem de programação incorporada no *Microsoft Excel* (Alexander e Kusleika, 2016; Jay, 2017).

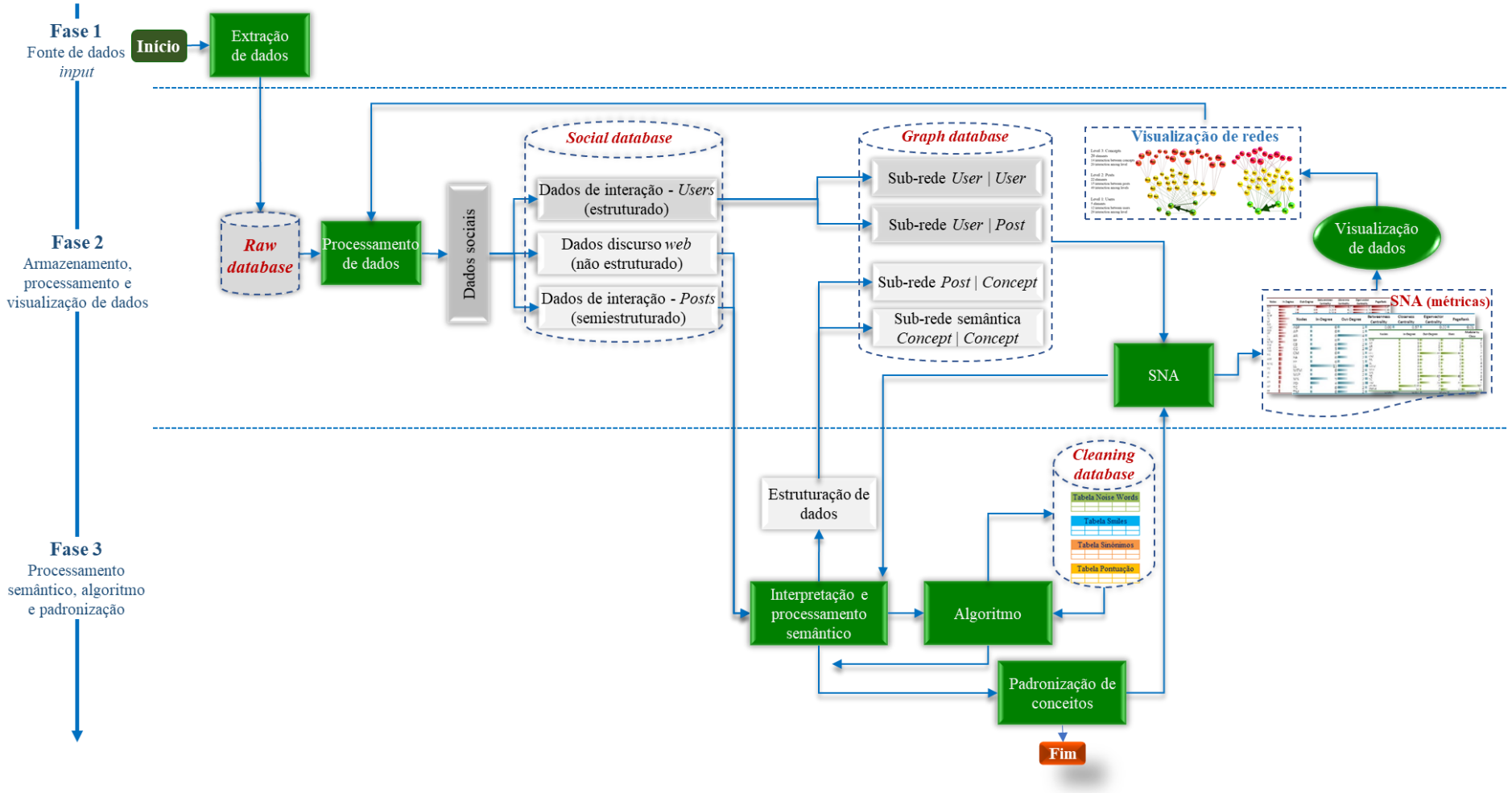
utilização deste algoritmo, foi a sua agilidade na manipulação dos dados e padronização dos conceitos visto que se trata de um processo recursivo e iterativo.

O *software Gephi* foi utilizado na segunda fase para visualizar, filtrar, manipular e exportar várias redes e para calcular as métricas mais importantes da SNA, pois apresenta um conjunto de características que permite realizar uma análise de rede através de processos analíticos/estatísticos e de técnicas de análise exploratória dos dados.

O modelo apresentado na Figura 21 traduz de forma global todo o fluxo de trabalho de análise dos dados das RSO para apoio à decisão. No entanto, existe também todo um conjunto de processos e componentes que são apresentados na Figura 22. As secções seguintes apresentam uma descrição detalhada do modelo, com particular ênfase para os processos que incorporam a segunda fase, visto ser a mais complexa e importante. A complexidade não está tanto no volume e diversidade dos dados, mas sim na dificuldade da sua estruturação e nos recursos disponíveis para os processar, que nem sempre estão acessíveis aos decisores.

Por uma questão de simplicidade, as fases do modelo encontram-se descritas de forma sequencial. Todavia, como o modelo é iterativo, na prática são necessários saltos de uma fase para outra. Por exemplo, na segunda fase, o processo SNA remete para o processo "Visualização de Dados" e este último remete para o processo "Processamento de Dados". Na terceira fase o processo Padronização de Conceitos remete para o passo SNA da segunda fase e assim sucessivamente.

Figura 22 - Componentes do modelo de análise de dados de RSO



5.1 Fase 1 - Extração de dados

O processo inicial do modelo envolve a recolha dos registos das RSO, seguindo-se o seu armazenamento e processamento. Nesta primeira fase do modelo, os dados referentes à estrutura da rede de utilizadores, trocas discursivas e respetivo conteúdo são extraídos de múltiplas fontes e armazenados de acordo com o seu formato e características.

As múltiplas fontes de dados a recolher permitem caracterizar e descrever as três entidades de análise, *utilizador* (*user*), *mensagem* (*post*), *conceito* (*concept*) e as respetivas ligações entre elas. O principal objetivo da extração é obter dados para interligar as entidades *user*, afetar um *post* a um *user* e, por sua vez, um *concept* a um *post*, para saber quem disse o quê. Assim, para além da caracterização da estrutura da rede de utilizadores, também será possível efetuar-se uma análise semântica das mensagens por eles trocadas.

As amostras de dados podem encontrar-se em formatos distintos, mas, independentemente da ferramenta utilizada, o importante é extrair dados das atividades dos utilizadores em torno do discurso *web* produzido *online*, ou seja, interações sociais e comunicação. Como referem Pang e Lee (2008), a extração é meramente um meio para atingir um objetivo.

Depois de reunidas todas as fontes necessárias para analisar as três entidades do discurso *web*, os dados são importados e estruturam-se na etapa "Processamento de Dados". Os dados armazenados na *social database* inferem-se com recurso à informação contida nos atributos das diferentes fontes de dados. Para o efeito, estabelecem-se relações entre os utilizadores e as mensagens que publicaram, quer *posts*, quer respetivos comentários. Estabelecem-se também relações entre as mensagens trocadas, isto é, entre o *post* que deu origem à troca discursiva e os comentários associados.

As RSO caracterizam-se por serem conjuntos de nós (entidades) que normalmente não possuem períodos de amostragem definidos. Caracterizam-se também pelo seu volume de dados, podendo implicar a necessidade de mais do que uma recolha. A frequência das extrações de dados a analisar depende, por isso, de vários fatores. Depende das limitações impostas não só pelas plataformas de RSO, mas também do volume de interações produzidas (volume de dados). Depende da capacidade do *hardware* e do *software* para recolher, armazenar e processar grandes volumes de dados. Depende ainda do fim a que se destinam os mesmos. Por exemplo, se os dados forem utilizados em cálculos diários, para o efeito são

necessárias extrações diárias. Se forem destinados a analisar acontecimentos isolados, serão necessárias extrações únicas.

Qualquer que seja a frequência com que os dados sejam extraídos, os mesmos armazenam-se, sem processamento, na base de dados referida no modelo como *raw database*. Todavia, intervalos de tempo relativamente curtos entre recolhas, garantem que os dados são o mais atuais possível (mais próximo do *real-time*) e, por isso, apresentam uma descrição mais precisa do comportamento dos utilizadores, bem como das suas trocas discursivas. Tais intervalos, definidos no modelo como *snapshots*, também permitem evitar os constrangimentos de demora que surgem na extração de grandes volumes de dados. A definição das unidades de tempo depende não só do objetivo da análise, mas também da quantidade de interações entre utilizadores, ou seja, do volume de dados das RSO.

Nas RSO a presença e a atividade dos nós e ligações mudam ao longo do tempo. Na realidade, os utilizadores criam e partilham mais informação do que criam novos contactos. Para além disso, um utilizador pode comentar uma mensagem, por exemplo, dois meses depois de ela ter sido criada por outra pessoa. Os eventos do mundo real apresentam dinâmicas temporais diferentes e a sua extração e visualização são, por si só, uma área de investigação onde vários métodos têm sido propostos, para resolver a diversidade de problemas existentes. O método proposto no modelo é a técnica designada de *parallelization*[44] que permite analisar as dinâmicas ocorridas nas RSO com recurso à visualização de sub-redes, criadas a partir dos *snapshots* extraídos em cada período de tempo. Este método facilita o processo de recolha, processamento e visualização do discurso *web* no seu todo.

A extração da informação produzida nas RSO é uma necessidade para o apoio à decisão em ambiente organizacional, mas, isolados, os dados são inúteis e sem sentido. Só após o seu processamento e transformação poderão adquirir sentido e significado.

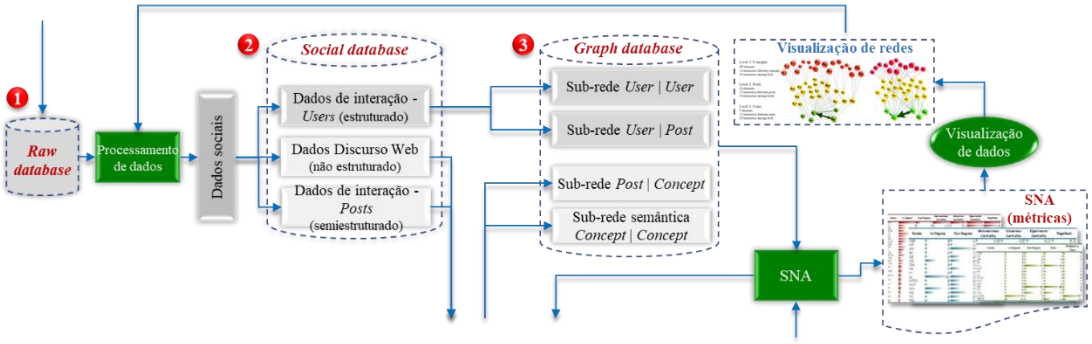
5.2 Fase 2 - Armazenamento, processamento e visualização de dados

A segunda fase do modelo é a mais complexa, visto que envolve toda a gestão e manuseamento dos dados, também eles complexos, para posterior utilização. Como mostra a Figura 23, é composta por várias etapas, cada uma com um propósito. Esta fase termina

[44] *Parallelization* é uma técnica de recolha e tratamento de informação que utiliza amostras independentes de dados, quando os recursos SI e TI são limitados ou porque a análise assim o exige (Khan et al., 2019).

com os dados estruturados, prontos para que sejam calculadas métricas da SNA e visualizadas as redes. Aparentemente, a Figura 23 mostra um processo sequencial e direto. No entanto, as várias etapas, sendo interligadas entre si, requerem interação com outras fases. A sua descrição apresenta-se na seguinte sequência: armazenamento de dados; processamento dos dados; cálculo de métricas da SNA e visualização dos dados.

Figura 23 - Fase 2: Armazenamento, processamento e visualização de dados



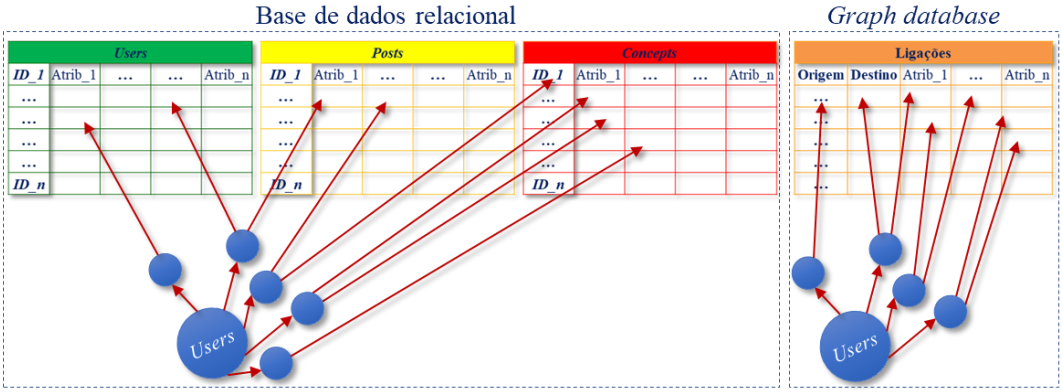
5.2.1 Armazenamento

A Fase 2 incorpora, ao longo de todo o processo, tecnologias de armazenamento de dados, onde se incluem duas bases de dados relacionais e uma não relacional. As duas primeiras bases de dados, assinaladas com o número 1 e número 2 na Figura 23 e designadas de *raw database* e *social database* armazenam, respetivamente, os dados fonte sem processamento e os dados processados e estruturados. A base de dados não relacional, assinalada com o número 3 na Figura 23, designada *graph database*, armazena os dados estruturados em formato grafo, para construção e visualização das redes finais.

A utilização de dois tipos de bases de dados diferentes serve os seguintes propósitos. As bases de dados relacionais, através de técnicas de *data-mining*, estruturam e processam as várias fontes de dados através de uma relação de um-para-muitos ($1 \times n$). Contudo, a sua estrutura de dados não permite modelar ou armazenar relacionamentos complexos do tipo muitos-para-muitos ($n \times n$). Assim, utiliza-se a *graph database* para modelar as ligações desses relacionamentos e otimizar o acesso aos dados. Enquanto que nas bases de dados relacionais as relações entre os registos são inferidas no momento da consulta, nas *graph databases* essas relações são diretamente armazenadas. Esta ação, qualquer que seja o volume de dados, torna o modelo mais eficiente na gestão e consulta de dados interligados.

Os dados na *graph database* estruturam-se recorrendo a um modelo de dados de origem e destino, isto é, armazenam-se e agregam-se os dados dos nós e as ligações entre eles. A Figura 24 exemplifica as diferenças de relações nos dois tipos de bases de dados.

Figura 24 - Tabelas relacionais *vs graph database*



Pode considerar-se esta segunda fase como o pilar do modelo, pois se todos os dados estiverem bem estruturados e sem erros, toda a informação extraída das bases de dados é fidedigna e satisfaz os requisitos para ela definidos.

No modelo, a função da *raw database* é armazenar dados estáticos em que os registos mantêm todos os atributos originais, sem alterações. A *raw database* recebe como *input* os dados tal e qual como foram recolhidos na fase anterior (extração de dados). Guardar os dados tal e qual como foram recolhidos tem algumas vantagens. Entre elas destacam-se: a possibilidade de explorar os dados[45] antes do início do processamento propriamente dito (esta técnica permite caracterizar os dados com recurso às métricas e técnicas de visualização da SNA); a possibilidade de voltar atrás no processo de processamento, sem grande esforço e sem perda de informação; destaca-se ainda a recuperação de informação[46] ou utilização de outras abordagens de processamento.

Relativamente à tecnologia de armazenamento *social database*, a mesma projetou-se para armazenar as três entidades do discurso *web* (*user*, *post* e *concept*). Nesta base de dados incluíram-se várias tabelas que permitem relacionar facilmente os dados das três entidades e depois aplicar técnicas do domínio do *data-mining*. Esta base de dados

[45] Exploração dos dados, tradução de *data understanding*, que é o processo de explorar e interpretar dados (Bassan e Sarkar, 2014; Moreira et al., 2019; Provost e Fawcett, 2013).

[46] Recuperação de informação, tradução de *data retrieval*, que é o processo de identificação e extração de informação de documentos guardados em bases de dados (Croft et al., 2015).

relacional é composta por três tabelas onde os dados estão estruturados de acordo com as suas características, isto é, se são dados estruturados, semiestruturados ou não estruturados.

Assim, na primeira tabela mencionada na figura como "Dados de interação - *Users"*, armazenam-se os dados das interações sociais entre utilizadores. Estes dados vêm, à partida, estruturados pelos *softwares* de extração de dados, não sendo necessários grandes processamentos para a sua utilização. Os mesmos podem ser diretamente armazenados na *graph database*, bastando identificar e uniformizar os atributos considerados importantes e relevantes para a análise.

A segunda tabela, mencionada na figura como "Dados de interação - *Posts*", tem como finalidade armazenar os dados das trocas discursivas do discurso *web*. São os dados da interação entre utilizadores e ligações entre as entidades *users* e *posts*. Estes dados caracterizam-se por serem semiestruturados, porque é necessário o seu processamento para identificar quem respondeu a quem e associar cada mensagem ao seu respetivo autor. Os *softwares* de extração de dados só extraem ficheiros com a informação da ligação *user* e *post* inicial, logo é necessário inferir as outras ligações em falta e necessárias para representar o discurso *web*. Todas as ligações em falta, entre as entidades *user* e *post*, constroem-se a partir da fonte de dados que contém os comentários dos utilizadores. Assim, são criadas as ligações entre o *post* inicial e respetiva resposta (*comment*), ou seja, ligação entre *post* e *post*. São também criadas as ligações entre os utilizadores envolvidos, *user* e *user*, para identificar quem respondeu a quem. O processamento envolve a importação dos dados dos ficheiros fonte e a sua estruturação na respetiva tabela da *social database*, como já referido.

Após este processamento estes dados transformam-se e servem de *input* para a *graph database*. Cada *post* (ou comentário de um *post*) é representado na base de dados como um único elemento (nó da rede), independentemente do tipo de informação que contém (quer seja texto, quer sejam outros caracteres, quer tenha muito quer tenha pouco texto). Deste modo identifica-se o utilizador que comentou um *post*, ou seja, que estabeleceu uma interação com o dono desse *post* e qual o *post* que foi comentado e com isto estabelecemos uma relação entre ambos os *posts* (*post* comentado e *post* resposta).

A terceira tabela, mencionada na figura como "Dados Discurso *web*", armazena os dados contidos em cada *post*. Na literatura, designam-se de dados não estruturados porque não estão nem padronizados, nem processados para uma utilização imediata. Para o efeito é necessário dividir cada *post* em tantos pedaços quantos os conceitos nele existentes. Estes

dados não são facilmente interpretáveis, uma vez que os textos não estão estruturados, os conceitos não estão uniformizados e podem estar escritos em diversos idiomas.

Antes do seu armazenamento na *graph database* é necessário estruturar, filtrar, limpar e uniformizar estes dados. A estruturação dos dados não estruturados, ou seja, o texto contido nos *posts*, executa-se com recurso a técnicas do PLN, recuperação de informação e *text-mining* na terceira fase do modelo. A partir da fonte de dados que contém os *posts* e da fonte de dados que contém os comentários, extraem-se todas as trocas discursivas (*posts*) e constroem-se todas as ligações entre as entidades *concept* e *post*, e *concept* e *concept*.

Relativamente à *graph database*, esta utiliza-se no modelo como ferramenta de armazenamento dos dados em formato grafo. Como já referido, um grafo é um conjunto de nós de uma ou várias entidades, que representa a forma como estas se relacionam entre si. Todos os relacionamentos sociais (entre utilizadores) e as ligações entre as entidades *post* e *concept* representam-se e armazenam-se na *graph database*.

A estruturação dos dados na *graph database* executa-se com recurso a técnicas de *data-mining* e *text-mining*. O *data-mining* utiliza-se no modelo na estruturação das entidades *user* e *post*. Já o *text-mining* utiliza-se para processamento semântico da entidade *concept* e para procurar informação nas fontes de dados não estruturadas, mais concretamente em dados em linguagem natural. Desta forma obtêm-se dados estruturados das ligações entre nó de origem e nó de destino para todas as entidades que a seguir se descrevem:

- *User* | *User* - Pares de nós da entidade *user* que estão ligados uns aos outros por uma relação de interação entre utilizadores. Os dados a recolher na primeira fase não têm todas as interações entre utilizadores, sendo necessário criar as ligações entre o utilizador que comentou e o utilizador do *post* inicial. Para o efeito, utilizando técnicas de *data-mining*, obtêm-se as ligações referentes à interação entre esses utilizadores;

- *User* | *Post* - Pares de nós das entidades *user* e *post* que ligam as duas entidades por uma relação de escrita. Os dados a recolher na primeira fase só contêm as ligações entre utilizador e o *post* inicial que deu origem à troca discursiva. Não existem dados com a relação entre o *post* (comentário) e o utilizador que o criou. É necessário inferir a relação entre o utilizador e o comentário que escreveu como resposta a outro *post*;

- *Post* | *Post* - Pares de nós da entidade *post* que estão ligados uns aos outros por uma relação de troca discursiva entre utilizadores, isto é, resposta dada a um *post*. As ligações entre as trocas discursivas também são inferidas das respostas (comentários) às mensagens iniciais;
- *Concept* | *Post* - Pares de nós das entidades *concept* e *post* que estão ligados numa relação em que o conceito está contido na mensagem. Estes dados obtêm-se na fase 3 do modelo, processamento semântico, com auxílio de um algoritmo;
- *Concept* | *Concept* - Pares de nós da entidade *concept* que estão ligados uns aos outros por uma relação de semântica e de sintaxe. As ligações entre conceitos, para produzir resumos das mensagens, constroem-se na terceira fase do modelo com técnicas de *text-mining* e auxílio do algoritmo.

O contexto organizacional é todo ele um mundo interligado onde não há informação isolada, mas sim interligada entre si. Só as bases de dados, tais como as *graph database,* que aceitam estes relacionamentos como característica central do seu modelo de dados, são capazes de armazenar, processar e analisar as ligações do discurso *web* de forma eficiente.

A estrutura da *graph database* definida no modelo, para além do par de valores origem e destino das entidades interligadas entre si (nós) pode também ter outros atributos. Os atributos contêm informação que caracteriza as entidades origem e destino, o tipo de rede, etc. Para além de contextualizar os nós e as ligações, os atributos permitem anexar metadados com informação sobre critérios ou restrições a determinados nós ou ligações. De referir que nas *graph database* há uma regra de consistência básica: não há ligações partidas porque uma relação tem sempre um nó inicial e um final. Tendo isso em conta, não se pode eliminar um nó sem também excluir as ligações a ele associadas.

A Figura 25 exemplifica o modelo de dados implementado da *graph database* para armazenar a informação de um dos estudos de caso. A *graph database*, para além de uma lista com todos os nós, contém uma lista com os registos das ligações que representam os relacionamentos com outros nós da rede (sejam ou não inter ou intra entidade). Esses registos de ligações estão organizados por um sentido que identifica a direção do nó (origem, destino) e, como referido anteriormente, contêm outros atributos adicionais. Os dados finais da *graph database*, armazenados em duas tabelas, uma de nós (*nodes*) e outra de ligações (*edges*),

constroem-se com recurso à base de dados relacional *social database* e a técnicas de *data-mining* e *text-mining*.

Figura 25 - Exemplificação do modelo de dados implementado

Nós (*nodes*)		
Atributo	**Descrição**	**Observações**
Id_Node	ID único de cada nó	Chave primária do nó. Relação 1 × n (um-para-muitos)
Entity_Type	*User, Post, Concept*	*User*: utilizador, *Post*: mensagem, *Concept*: conceito
Post	1º *post*, início da troca discursiva	
Owner_Post	1º *user* que iniciou troca discursiva	
Comment_Post	Resposta ao 1º *post* da troca discursiva	
Owner_Comment_Post	*User* que respondeu a um *post*	
…	Outros atributos	
Atrib_n		

Ligações (*edges*)		
Atributo	**Descrição**	**Observações**
Source	Nó de origem da ligação	Par de atributos obrigatórios no modelo *graph database*. Estabelecem relação n × n (muitos-para-muitos).
Target	Nó de destino da ligação	
Network_Type	Directed	
Source_Type	*User, Post, Concept*	Atributos que caracterizam a entidade a que pertencem os nós.
Target_Type	*User, Post, Concept*	
Post	1º *post*, início da troca discursiva	Atributos que permitem interligar todas as entidades para saber quem disse o quê.
Owner_Post	*User* que iniciou troca discursiva	
Comment_Post	Resposta ao 1º *post* da troca discursiva	
Owner_Comment_Post	*User* que respondeu a um *post*	
Edge_Type	*User* | *User, User* | *Post, Post* | *Post, Concept* | *Post, Concept* | *Concept*	Atributo que caracteriza as ligações inter e intra entidade.
Semantic_Network_Type	Caracterização da rede semântica, do tipo resumo ou palavras-chave	Atributo que caracteriza os tipos de redes semânticas.
…	Outros atributos	
Atrib_n		

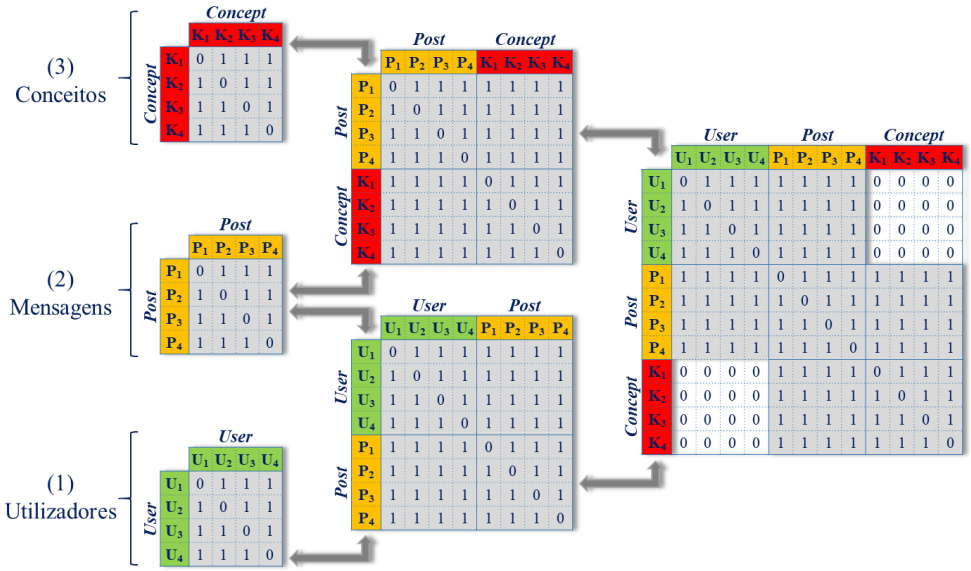
Por norma, os nós utilizam-se para representar as entidades mas, dependendo do tipo de análise, as ligações também podem ser utilizadas para o efeito. Por esse motivo, na tabela de ligações, é criado um atributo que caracteriza o tipo de ligação. O atributo assume valor *User* | *Post* quando caracteriza as ligações entre as entidades *user* e *post*, valor *Concept* | *Post* quando caracteriza as ligações entre o *concept* e o *post*, etc. A função deste atributo é agrupar as ligações de acordo com o seu tipo em conjuntos distintos. Caracterizando o tipo de ligação entre nós, é possível identificar se a relação é dentro do mesmo nível de análise (intra-entidade) ou entre níveis diferentes (inter-entidade). Para além dos atributos que caracterizam os nós e as ligações, existem outros que permitem estabelecer uma relação entre os nós do terceiro nível (*concept*) e os do primeiro nível (*user*), sem passar pelo segundo nível. Para saber qual (ou quais) os utilizadores que utilizaram um determinado conceito, não é preciso saber a que *posts* pertencem. Com a utilização de atributos consegue-se identificar quem escreveu o quê e, caso necessário, em que *post*.

A função dos atributos, para além de servir para agrupar os nós em subconjuntos, permite também, caso necessário, trabalhar individualmente com eles. Assim, em vez de se

trabalhar com todos os dados da RSO extraída, trabalha-se só com uma amostra de dados. Isto permite que as consultas e a criação de redes sejam mais rápidas e eficientes quando estamos perante um grande volume de dados. Em particular para a análise semântica, que é recursiva, é mais eficiente trabalhar só com a sub-rede de conceitos na fase de *data cleaning* para identificar conceitos irrelevantes a eliminar ou corrigir.

No modelo, a abordagem utilizada para analisar em simultâneo os três níveis da rede foi a transformação de redes *two-mode* numa rede *one-mode*. Matematicamente, cada entidade em análise representa-se por uma matriz. Este método funciona, ligando dois conjuntos de dados um ao outro, através de nós comuns a ambos. O resultado é uma matriz única com todas as entidades em análise (*user, post* e *concept*) representada na Figura 26.

Figura 26 - Exemplificação da transformação dos dados para a matriz *two-mode*

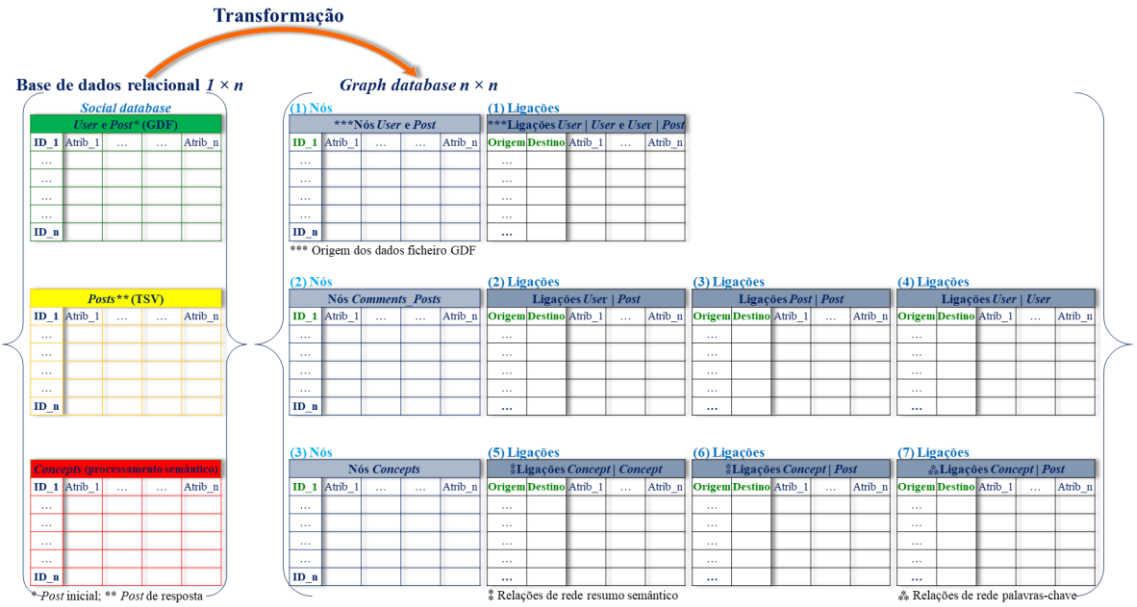


5.2.2 Transformação da *social database* para *graph database*

Como já referido, nos dados a recolher não existem todas as relações entre entidades, necessárias para uma análise completa da estrutura da rede e dos conteúdos produzidos pelos utilizadores. As bases de dados relacionais incluíram-se no modelo para que fosse possível inferir dados e estruturá-los como base para construir relações do tipo $n \times n$ (muitos-para-muitos). Porque nem todas são fornecidas pelos dados extraídos pelos *softwares* de extração, é necessário um processo de transformação dos dados do modelo relacional para o modelo *graph database*. A Figura 27 exemplifica o processo de transformação dos dados, implementado em folhas de cálculo, de um modelo para outro.

O tipo de relacionamentos $n \times n$, ao contrário dos de $1 \times n$ em bases de dados relacionais, exige inferir ligações entre entidades utilizando identificadores únicos (*IDs*). Depois de construir as ligações em falta entre os nós das várias entidades, é possível obter dados estruturados que descrevem não só a estrutura das RSO, mas também análises semânticas das trocas discursivas.

Figura 27 - Transformação de dados do modelo relacional para *graph database*



A transformação dos dados das tabelas relacionais para as tabelas com relacionamentos do tipo $n \times n$, executa-se com recurso a três colunas definidas para o efeito. Cada coluna, ou atributo, contém informação com metadados que permitem inferir através dos *IDs* as ligações em falta. O primeiro e o segundo atributos caracterizam a entidade de origem e de destino de cada nó. Este par de atributos permite construir as ligações entre as várias entidades. O terceiro atributo da tabela caracteriza o tipo de ligação. O resultado final, após a transformação dos dados, é uma tabela com todos os nós das três entidades e uma tabela com todas as ligações. Isto é, dados estruturados em formato de grafo.

A transformação dos dados tem como objetivo integrar os dados na *graph database*, depois de processados na base de dados relacional criada em *Excel*. Somente depois de todos os dados estarem estruturados, quer seja porque vieram diretamente das fontes de dados estruturados, quer seja porque passaram pelo processamento semântico na segunda fase do modelo, é possível armazená-los na *graph database* para construir dois tipos de redes: uma

para criar redes semânticas com o resumo das trocas discursivas e outra para criar redes semânticas com palavras-chave.

De referir que, para construção da rede semântica de palavras-chave, não é necessário o subconjunto de dados que estabelece a ligação entre conceitos (*concept* | *concept*). Esta ligação só é necessária para a criação dos resumos semânticos pois identifica o conceito que está imediatamente antes e o que está depois, o que permite reconstruir o texto das mensagens.

Após a transformação, limpeza e estruturação dos dados, estes armazenam-se na *graph database* para que seja possível realizar a sua análise e manipulação. Só depois disso se criam resumos visuais definitivos, se recalculam métricas da SNA e se avalia se os resultados são úteis ou não para o apoio à decisão.

5.2.3 Visualização dos dados e métricas SNA

Podemos considerar que a representação gráfica de uma rede, que engloba todos os utilizadores, mensagens e conceitos é um resumo do discurso produzido nas RSO. A técnica de visualização utilizada no modelo permite uma interpretação mais intuitiva dos dados dos diferentes níveis de análise, ajudando a acelerar não apenas o seu processamento, mas também a análise semântica descrita no subcapítulo 5.3.

No modelo, a análise visual utiliza-se quer para auxiliar na eliminação dos dados irrelevantes da rede de utilizadores, quer para identificar nós importantes na rede semântica. Isso é feito utilizando grafos distintos e/ou interligados, que incluem as redes de utilizadores, de mensagens e as redes semânticas. A interligação dos *users* aos seus *posts* permite estruturar um resumo do discurso, onde cada entidade representa uma sub-rede. As representações gráficas resumem o discurso *web*, mostram uma visão geral do *post* original, permitem explorar a estrutura social das RSO e a semântica contida nas mensagens. Os nós de cada sub-rede diferenciam-se, atribuindo pesos diferentes para identificar e selecionar os elementos mais relevantes da rede semântica.

No modelo, a visualização para análise dos dados semânticos é um processo recorrente e iterativo, onde o envolvimento humano é essencial para uma análise sistemática dos resultados obtidos, bem como para verificar se são necessários ajustes. Deste modo, o conhecimento do contexto organizacional capta-se manualmente, transfere-se e armazena-se na *cleaning database*, para ser utilizado de forma automática até afinar todo o processo de

limpeza e uniformização dos dados. Após identificação dos dados considerados irrelevantes, como explicado no subcapítulo 5.3.1, a *cleaning database* configura-se para os descartar num segundo processamento.

É a partir da estrutura da rede, composta pelas três entidades, *user*, *post* e *concept*, que é possível estabelecer relações entre o processo de comunicação e o respetivo discurso *web* como um todo, bem como as suas aplicações no apoio à decisão. Por um lado, a SNA pode ajudar a obter informação útil sobre a estrutura da rede e padrões de relacionamento tais como quem fala com quem, que conteúdo foi transmitido e quem está ligado a quem. Por outro, um conceito é uma ideia única representada por uma ou mais palavras, sendo possível utilizar a metodologia SNA para processar e representar a rede de conceitos obtida.

5.2.3.1 SNA para apoio ao processo de estruturação dos dados sociais

Após a extração, com o apoio da SNA, métricas e visualização, os três níveis do discurso *web* encadeiam-se de forma coerente para facilitar a sua estruturação. Os dados, utilizadores (*user*) e trocas discursivas (*post*), são desta forma interpretados, armazenados e caracterizados para definir de forma cautelosa o que cada rede representa. Para o efeito recorre-se à SNA que simplifica a compreensão e interpretação da informação, tornando o processo de *data-mining* e extração de conhecimento mais simples. Como estamos perante dados sociais em linguagem natural, é importante explorá-los de forma a recuperar informação útil para o apoio à decisão organizacional.

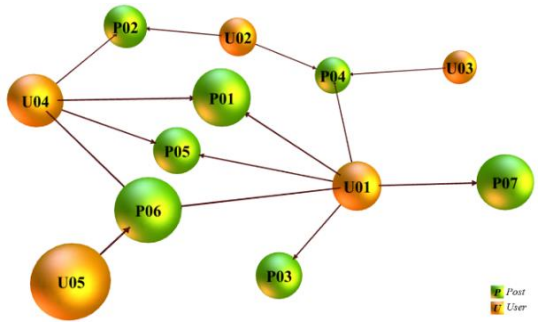
Explorar os dados permite identificar e estruturar as relações que existem, entre as diferentes fontes de dados e os seus atributos-chave e identificar a necessidade ou não de transformações. Importa referir que a identificação do tipo de entidades a que pertencem os nós, o tipo de ligações entre elas e se existem ou não nos dados, faz-se no início do processo com recurso à exploração dos dados. Nessa etapa identificam-se também os atributos necessário para caracterizar quer os nós, quer as ligações entre eles. Permite ainda compreender o contexto dos dados e definir qual a sua utilização, bem como identificar características dos textos, tais como informalidade e extensão.

Tendo como objetivo a exploração e interpretação, os dados extraídos, isto é, sem qualquer processamento inicial, visualizam-se no *software Gephi* para identificar nós irrelevantes, que não interessem para a análise. A interpretação executa-se utilizando técnicas matemáticas e de visualização, para avaliar a qualidade dos dados. Ilustra-se na

Figura 28 a representação gráfica dos dados estruturados recolhidos na primeira fase do modelo, sendo mais fácil representar e interpretar as entidades *user* e *post* através da visualização da rede. Para uma rápida identificação visual, os nós têm uma cor associada de acordo com o tipo de entidade. Os nós da rede são os utilizadores, a entidade *user* é representada a verde e a primeira mensagem da troca discursiva, uma entidade *post* a amarelo.

Com apoio da SNA, os dados encadeiam-se de forma coerente para facilitar o seu acesso e a sua análise. Através da SNA e utilização das métricas *in-degree* e *out-degree*, são extraídos dados relevantes (ou se quisermos são eliminados dados irrelevantes). São igualmente eliminados dados referentes a nós sem ligações, ou seja, eliminam-se nós (normalmente *posts*) com *in-degree* e *out-degree* iguais a zero. Estes nós sem ligação a outros, são *posts* que derivam de mensagens automáticas colocadas pelo próprio *Facebook*.

Figura 28 - Visualização gráfica das entidades: *user* e *post*



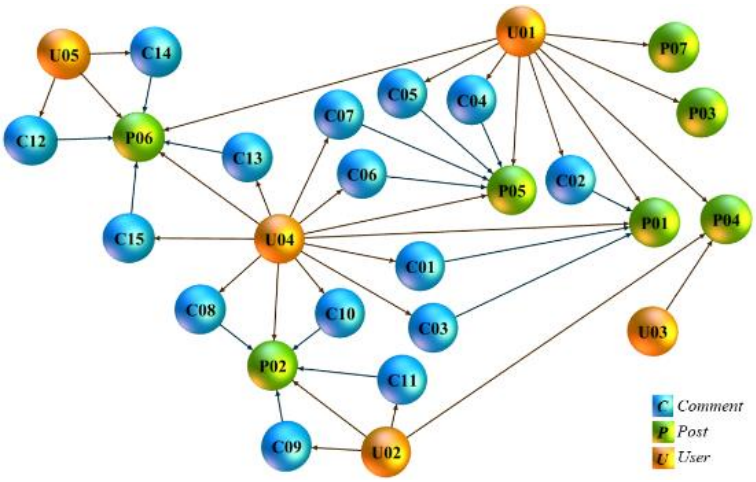
Após a limpeza dos dados irrelevantes, existentes no conjunto de dados com as entidades *user* e *post*, constrói-se uma nova rede onde são incluídos os comentários efetuados pelos utilizadores. Essa rede, exemplificada na Figura 29, representa a junção do conjuntos de dados da rede anterior (Figura 28) com os dados da atividade discursiva dos utilizadores, isto é, todos os comentários feitos a um *post*. Os dados representados na rede da Figura 29 são semiestruturados pois encontram-se em formatos e fontes de dados distintas. Na fase de processamento, os dois conjuntos de dados agregam-se para construir conjuntos de dados com as ligações, em falta, entre mensagem e comentário.

Quando os dados são interpretados (logo após a recolha e na fase de processamento), faz-se a distinção entre o que é um *post* e o que é um *comment* porque, apesar de ambos serem *mensagens*, encontram-se em fontes de dados distintas. Por esse motivo, na Figura 29 faz-se a distinção entre os *posts* (nós verdes) e os comentários (nós azuis). Por uma questão de simplificação e levando em consideração que quer um *post*, quer

um *comment* são um conteúdo textual, logo a mesma coisa, no modelo final uniformizou-se e ambos foram definidos por *post*.

Com auxílio da SNA, o discurso *web* estrutura-se e as trocas discursivas são encadeadas de forma coerente para facilitar o seu acesso e a sua análise. Através da visualização da rede apresentada na Figura 29, é possível identificar como cada interveniente se relaciona com os outros e como contribui com informação. A formalização e representação gráfica do discurso *web* permite identificar quem escreveu uma mensagem e a quem.

Figura 29 - Visualização gráfica das entidades: *user*, *post* e *comment*



Após a eliminação dos dados irrelevantes das entidades *user* e *post*, níveis de análise 1 e 2 respetivamente, passa-se à análise do conteúdo textual. As trocas discursivas, mensagens e os comentários (ambos representados no modelo pela entidade *post*) são transformados em redes de palavras. A extração dos conceitos (entidade *concept*) a partir dos *posts* executa-se com recurso ao *Excel*, para serem depois importados e visualizados com o *Gephi*.

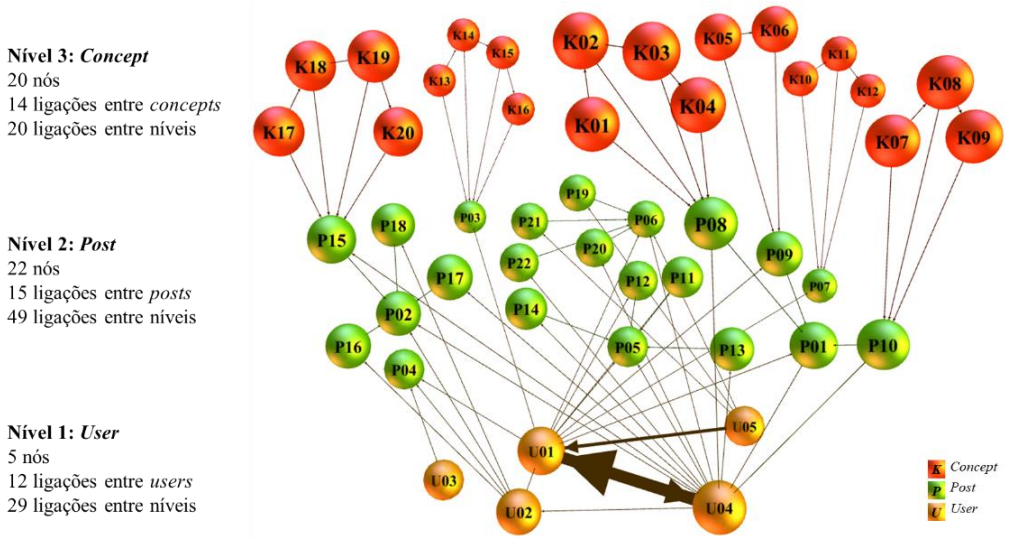
A SNA simplifica a compreensão e interpretação dos dados semânticos, terceira fase do modelo, tornando o processo de *text-mining* mais simples para a padronização dos conceitos que alimentam a *cleaning database*. Como já referido, o texto é transformado em dados interpretáveis pela SNA, partindo cada *post* em conceitos individuais. Ou seja, em múltiplas unidades individuais que definem cada conceito como sendo um nó da rede semântica. Esta transformação permite estruturar os dados textuais para construção das redes com os três níveis de análise. Os dados não estruturados são interpretados e explorados através de um processo recorrente e iterativo, com recurso a *data-mining* e *text-mining*

através da SNA. Isto é, através do cálculo das métricas da SNA e visualização são identificados os conceitos irrelevantes a alterar, para atualizar a *cleaning database*.

Levando em consideração que o discurso informal, com foco nos diálogos, pode ser convertido em dados semânticos e que a visualização auxilia nos processos de extração de informação, as entidades *concept* e *post* interligam-se com recurso a dois métodos. O primeiro, permite obter um resumo de cada troca discursiva, enquanto que o segundo constrói uma rede de conceitos.

Para resumir o conteúdo dos *posts* (primeiro método), é necessário identificar palavras vizinhas. O resultado desse processo é uma sub-rede semântica, na qual cada conceito é definido como um nó da entidade *concept*, independentemente de este existir ou não noutro *post*. Deste modo, resume-se o conteúdo de cada *post* através de uma sub-rede semântica, como representado na Figura 30. A sub-rede semântica constrói-se estabelecendo uma relação entre os pares de *concept* contidos no *post* e representados na figura a vermelho. Para o efeito, estabelece-se uma relação entre o $concept_1$ e o $concept_2$, entre o $concept_2$ e o $concept_3$, e assim sucessivamente.

Figura 30 - Visualização gráfica do discurso *web*: resumo semântico



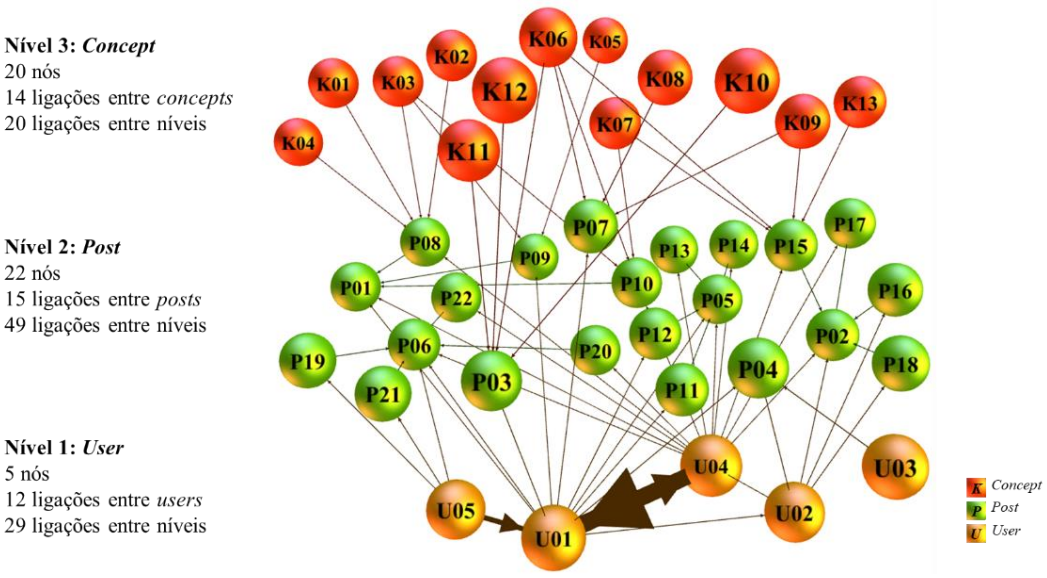
A identificação e uniformização dos conceitos da rede executa-se não só com recurso a algoritmos criados para o efeito (como explicado no subcapítulo 5.3.2), mas também com recurso à SNA. Com recurso às métricas e à visualização da SNA, os conteúdos dos *posts* interpretam-se e exploram-se para verificar se são necessários ajustes. Nesta etapa, o envolvimento humano é essencial para uma análise sistemática e interativa dos resultados semânticos obtidos. A ideia subjacente à utilização de técnicas de visualização para PLN é

simples. A intervenção humana justifica-se, pois a utilização de técnicas de processamento semântico para atualização da *cleaning database* dependem da informação encontrada.

A rede da Figura 30 ilustra como cada utilizador (*user*) contribuiu com mensagens (*post*), estabelecendo uma troca discursiva que representa uma ligação entre eles, e resume o conteúdo (*concept*) das mensagens. A sua visualização representa um resumo de comentários e/ou opiniões expressas nas trocas discursivas, que é importante para perceber se a informação em análise deve ser ou não considerada. Isto é, permite identificar, de forma visual, se a informação nela contida será apropriada para o apoio à decisão. A visualização da rede, com as três entidades, num grafo, permite ver padrões que antes não eram evidentes. Conjugando a visualização com as métricas da SNA, esses padrões sobressaem e permitem identificar se os conceitos são ou não irrelevantes e/ou inúteis.

Enquanto que na primeira etapa (rede da Figura 30) cada *post* tem associada uma sub-rede semântica, na segunda qualquer *post* pode estar ligado à rede semântica global, como mostra a Figura 31. Com esta etapa, as trocas discursivas entre utilizadores estruturam-se de forma a criar redes semânticas de palavras-chave. Para o efeito, cada conceito (*concept)* é um nó com um *ID* único, ou seja, só ocorre uma vez dentro da rede semântica global. O objetivo é detetar palavras-chave, visto que uma questão relevante no *text-mining* para análise semântica é a capacidade de procurarmos nos dados uma informação específica. No modelo, a extração de palavras-chave é uma técnica importante para identificação dos conceitos mais utilizados e relevantes nos *posts*.

Figura 31 - Visualização gráfica do discurso *web*: palavras-chave



As métricas da SNA permitem modelar os nós da sub-rede semântica e muitas das suas características analisam-se recorrendo à manipulação e visualização de redes. Como ilustrado na Figura 31, a visualização da rede auxilia, em testes exploratórios, o processo de identificação de conceitos com o mesmo significado semântico. Por exemplo, "caixa" e "cx" têm o mesmo significado semântico, logo são padronizados para "caixa". Os resultados são, assim, utilizados para analisar e ajustar os dados de acordo com as necessidades. Depois de identificados e padronizados os conceitos, é atribuído o mesmo *ID* replicando as ligações da palavra-chave para o(s) *posts*.

As redes da Figura 30 e da Figura 31, agregam não só a interação entre utilizadores e as suas trocas discursivas, mas também redes semânticas. Nestas redes a espessura das linhas, que representa as ligações entre entidades, codifica o número de vezes que um utilizador comentou ou gostou do *post* de outro utilizador. A sua visualização simplifica o processo de captar e interpretar o fluxo de informação com base nas características dos dados. Para além disso, a visualização dos dados num grafo, com as redes semânticas de resumos e de palavras-chave, permite identificar eventuais dados irrelevantes no conteúdo dos *posts*. Aos nós das redes apresentadas nas figuras, foram atribuídas cores para uma melhor identificação visual dos três níveis. As entidades *user*, *post* e *concept* representam-se pelas cores verde, amarelo e vermelho, respetivamente. Visualmente, consegue-se identificar os nós que fazem parte de cada nível de análise.

5.2.3.2 SNA para apoio à decisão

Os dados finais criam por si só um processo de apoio à decisão, onde o decisor participa antes de desencadear uma tomada de decisão. A aplicação da SNA para análise das interações entre utilizadores e conteúdos publicados pode melhorar esse processo e a tomada de decisão pode ser sustentada com dados reais obtidos no momento adequado. Os resultados obtidos permitem entender como o discurso *web* pode ser utilizado nas várias fases da decisão, para a apoiar e influenciar, bem como explicar como as comunicações em tempo real, também a partir das RSO, alteram a decisão.

Os dados obtidos através do modelo proposto consultam-se e utilizam-se para apoio à decisão através de *software* adequado, tal como o *Gephi* e o *Excel*. Os dados finais produzidos podem ser representados visualmente em sub-redes, rede de interações entre utilizadores (*user* | *user*), ligações entre *posts* e conceitos e ligações entre pares de conceitos. A visualização foi incorporada no modelo como técnica para identificar informação

relevante para apoio à decisão, pois facilita a interpretação e extração de padrões e significados de representações visuais. Em contexto organizacional, pode dizer-se que a visualização é o caminho mais rápido para comunicar e interpretar dados para apoio à decisão.

As métricas da SNA, por norma, utilizam-se para quantificar interações entre utilizadores, contabilizar comentários de uma mensagem, medir fluxos de informação, entre outros. Podem ser utilizadas na apresentação dessa informação para facilitar a avaliação de alternativas, através de classificação e/ou identificação da posição de um nó. Também podem ser utilizadas para efetuar cálculos. Por exemplo, o significado semântico das palavras-chave e o número de vezes que foram usadas (*out-degree*) utilizam-se para realizar cálculos ou estimativas. A métrica *out-degree* indica quantas vezes uma palavra-chave foi utilizada ao todo pelos utilizadores, sendo aplicada no modelo para os casos em que o grafo representa algo como um resultado ou quantificação de uma situação. Exemplificando, para calcular quantidades encomendadas de um produto, pode utilizar-se o *out-degree* pois a rede representa uma situação estática e interessa o nó que tem o maior valor.

Os dados finais traduzem a informação para apoio à tomada de decisão de forma compacta e visualmente atraente. O modelo utiliza os métodos formais da SNA, matrizes e grafos, para representar as descrições dos padrões das interações sociais das RSO de forma compacta e sistematizada. Para além disso, as regras e as convenções dos métodos matemáticos e das técnicas de representação gráfica auxiliam na identificação de situações não identificáveis nos dados.

5.3 Fase 3 - Processamento semântico, algoritmo e padronização

A terceira fase do modelo envolve as etapas de processamento semântico dos dados lexicais das mensagens recolhidas. A implementação de um sistema computacional capaz de realizar tarefas que envolvam dados *online* implica especificar não só a estrutura dos dados, mas também o seu contexto específico. Assim, a fase de processamento semântico, requer a extração de todos os conceitos existentes nas mensagens e posterior tratamento para utilização final, num processo recursivo (o que significa que, até à obtenção final dos dados que alimentam a *graph database*, pode ter de ser executado diversas vezes).

Esta fase do modelo é composta por várias partes, nomeadamente a estruturação de dados; a interpretação e processamento semântico; o algoritmo para auxiliar o

processamento dos mesmos; uma *cleaning database* que auxilia nas etapas de limpeza dos dados, visto que os dados sociais podem conter informação que não é relevante e/ou é considerada lixo; de dados padronizados para construção de redes semânticas. Esta fase em que os dados semânticos não estruturados são convertidos em dados estruturados termina quando todos estão padronizados e estruturados, prontos para serem armazenados na *graph database*. Todo o processo é iterativo onde os passos (desta fase) dependem da fase anterior do modelo.

A *web* semântica tem por base a ideia de que os dados já estão definidos e estruturados na *web* de forma a que possam ser utilizados por pessoas e processados por máquinas. No entanto, ainda estamos longe disso porque é impossível atualizar a *web* e criar ontologias para todo o tipo de contextos organizacionais existentes. É impossível por dois motivos. Primeiro, a tecnologia ainda não tem capacidade de resposta para processar e representar computacionalmente a quantidade de dados produzidos nas RSO. Segundo, porque poderemos estar a falar de contextos tão extremos como analisar dados de uma empresa que vende morangos, ou de uma empresa de aviação. A linguagem utilizada numa é completamente diferente da outra, não existindo, assim, padrões (ontologias) já definidos e normalizados dentro da *web* semântica que possam ser utilizados pela maior parte das empresas. Portanto, enquanto a *web* semântica ainda não está totalmente implementada, é necessário criar mecanismos, como as *cleaning databases*, com dados específicos de cada contexto, para apoiar o PLN.

No modelo, a *cleaning database* é uma estrutura utilizada para representar e/ou classificar conceitos (texto) que possam ser utilizados por máquinas e que respondam de forma expedita à realidade de cada empresa. Nesse sentido, a *cleaning database* pode ser comparada a uma ontologia que descreve a estrutura e a semântica dos dados, utilizando padrões, auxiliando o seu processamento de forma automática.

Partindo da premissa de que qualquer contexto de decisão tem a sua própria ontologia, porque atribui significado aos símbolos que utiliza de acordo com a visão particular do seu mundo, o modelo contempla uma ontologia local. A mesma justifica-se porque a heterogeneidade semântica que existe entre contextos organizacionais é um problema para a utilização de padrões.

Como referido, cada contexto de tomada de decisão é diferente e tem as suas especificidades. Esse facto impossibilita que ontologias criadas no âmbito da *web* semântica sejam facilmente utilizadas em ambientes organizacionais multidisciplinares e heterogéneos. As ontologias locais, de um contexto específico e independente de decisão, representam o seu

ambiente de ação bem como o ponto de vista dos seus utilizadores. Por isso, criar mecanismos de representação de dados para cada contexto organizacional significa incorporar e desenvolver modelos que sejam iterativos e que tenham uma base de conhecimento desenvolvido no contexto específico.

A principal aplicação da *cleaning database* está associada à capacidade de o modelo interpretar o conteúdo disponível na *web* social para o conseguir perceber de forma distinta. Ou seja, (por exemplo, a palavra "manga" é uma fruta ou é parte de uma peça de vestuário) o conteúdo é interpretado de acordo com o contexto onde é utilizado. A *cleaning database* é assim uma estrutura para especificar formalmente a semântica contida nos conteúdos, através de componentes que envolvem a criação e implementação de ontologias.

No processamento semântico, incorporado no modelo, a tarefa de limpeza dos dados é minimizada com a utilização de um algoritmo que suporta a padronização do conteúdo das mensagens. É verdade que já existem ferramentas e algoritmos para o efeito (por exemplo, o *Trifacta Wrangler*[47] ou o *Wordnet*), mas esta fase permite uma otimização sistemática e recursiva do sistema. Isto é, cria condições para que todo o processo de padronização faça sentido, para o caso de ser necessário realizar alterações de dados *a posteriori*. Permite também que a limpeza dos dados seja feita de acordo com o respetivo contexto. Desta forma, consegue-se um vocabulário estruturado, controlado e de uso comum (dados informais de cada contexto) na *cleaning database*.

Em cada iteração utilizam-se os resultados do algoritmo para alimentar a *cleaning database*, melhorar e reforçar o desempenho do processamento semântico na iteração seguinte (até afinar).

5.3.1 Cleaning database

A *cleaning database* é uma estrutura que guarda listas de conceitos definidos como padrão para posterior comparação com os dados semânticos das trocas discursivas. A falta de qualidade dos dados, para além de enviesar os resultados põe em causa toda a análise. Por isso, a limpeza dos dados é importante para garantir que os mesmos sejam o mais precisos possível, respondendo à realidade de cada contexto organizacional. Para além disso, uma atualização continua da *cleaning database* torna o sistema mais "inteligente" para

[47] *Trifacta Wrangler* é um serviço que ajuda a limpar e preparar dados com a maior rapidez e precisão possíveis. Acesso em 16 de dezembro de 2019, disponível no portal da *Trifacta*: https://www.trifacta.com/.

interpretação dos dados porque a adição de novos conceitos é essencial para que haja correspondência entre os dados mais antigos e os mais recentes.

A limpeza dos dados e a sua preparação para análise é uma tarefa laboriosa visto que, na fase inicial (construção de todo o processo) é demorada, trabalhosa e meticulosa, independentemente do caminho utilizado para a fazer. A abordagem definida nesta fase do modelo permite a integração de novos dados de forma contínua, a sua consulta e visualização para acelerar a compreensão humana. O processo de limpeza dos dados não é automático e depende de alguém que conheça o contexto da rede, para incorporar na *cleaning database* a linguagem aí utilizada, para que a seguir a máquina a possa utilizar de forma automática. O custo de um erro pode ser considerável, pois podemos estar a evidenciar valores sem sentido e errados.

A *cleaning database* incorporada no modelo funciona como uma ontologia que fornece ao sistema a interpretação e a padronização de conceitos, que explicam, de forma explícita e formal, a semântica de um contexto organizacional específico. A *cleaning database* melhora a interoperabilidade das diferentes formas de comunicação contidas na linguagem das RSO. Permite também que os conceitos padronizados possam ser partilhados e reutilizados de forma eficiente. Assim, a construção da *cleaning database* é uma mais valia para a partilha e a reutilização dos conceitos existentes num contexto organizacional.

Em contextos organizacionais específicos, a utilização de padrões desenvolvidos para outros contextos pode não considerar conceitos que são importantes, mas não se encontram padronizados. Comparar texto já padronizado noutro sistema (base de dados, ontologias, etc.) com o texto de uma mensagem das RSO e procurar uma correspondência exata, pode não produzir os resultados esperados.

Utilizar uma mistura, isto é, um modelo que incorpora padrões e o conhecimento extraído do contexto de decisão é o ideal. Ou seja, criar uma estrutura (*cleaning database*) incorporada no modelo, que é alimentada por padrões (predefinidos noutros sistemas) e pelos conceitos extraídos do contexto organizacional. Se, por um lado, a *cleaning database* é enriquecida com dados oriundos e já validados pela literatura, por outro enriquecemo-la com o conhecimento do contexto de decisão. Assim, foram estes os motivos que levaram à incorporação, no modelo, de uma *cleaning database*, alimentada com os conceitos encontrados em cada contexto de decisão. Para além de garantir uma maior fiabilidade dos dados, também permite um maior controlo dos mesmos ao permitir controlar o vocabulário característico de cada contexto organizacional. O modelo não fica dependente nem de fontes

externas (*thesaurus*, dicionários, ontologias, etc.), nem de *software* externo para interpretar os dados das trocas discursivas. Os critérios e restrições de eliminação e/ou correção dos dados definem-se de acordo com o contexto organizacional. Isto permite criar listas de classificação de conceitos mais precisas, para serem utilizadas pelo algoritmo criado para estruturar o texto das mensagens. Todos os dados são considerados, pelo que não estão sujeitos à sua pré-existência em dados predefinidos de fontes externas. Não há uma lista de palavras-chave predefinidas *a priori*. As palavras-chave são identificadas através das métricas da SNA (*degree* elevado = palavra-chave; *degree* baixo = *outlier*[48]).

As tabelas da *cleaning database* encontram-se organizadas de acordo com a sua função e de acordo com o tipo de processamento que se pretende fazer. Estas tabelas também são designadas na literatura como *bag-of-words*[49]. Esta base de dados contém quatro tabelas a que se deu o nome de *Smiles*, Pontuação, *Noise-words* e Sinónimos, como mostrado na Figura 32. As duas primeiras são simples quer na forma como são atualizadas, quer no seu conteúdo. Relativamente às outras duas, são tabelas mais complexas pois, para além de serem alimentadas com recursos existentes na literatura, são alimentadas manualmente à medida que são identificados os conceitos a eliminar ou a alterar. As tabelas têm atributos que caracterizam os significados dos conceitos, as relações existentes entre os vários significados e definições de palavras que, em conjunto, têm um significado.

Figura 32 - *Cleaning database*: exemplo de tabelas

Tabela de *Smiles* do *FaceBook*

Id	*Smile*	Significado	Uniformização
1	:-)	sorriso	:-)
2	:)	sorriso	:-)
3	:]	sorriso	:-)
4	=)	sorriso	:-)
5	:-(	triste	:-)
6	:(	triste	:-)
7	:[	triste	:-)
8	=(	triste	:-)
9	:O	surpreendido	:-)
10	:-o	surpreendido	:-)
11	:o	surpreendido	:-)
12	:-D	sorriso rasgado	:-)
13	=D	sorriso rasgado	:-)
14	:-P	língua de fora	:-)
15	:P	língua de fora	:-)
16	:-p	língua de fora	:-)
17	:p	língua de fora	:-)
18	=P	língua de fora	:-)
19	;-)	piscar de olho	:-)
20	;)	piscar de olho	:-)
21	:-*	beijo	:-)
...			

Tabela de *Stopwords/noise-words*

Id	*Stopword*	Tipo
1	a	Artigo
2	à	Artigo
3	aos	Artigo
4	um	Artigo
5	uma	Artigo
6	o	Artigo
7	os	Artigo
8	alguns	Artigo
9	pelo	Artigo
10	pela	Artigo
11	do	Artigo
12	das	Artigo
13	da	Artigo
14	apenas	Preposição
15	apesar	Preposição
16	para	Preposição
17	após	Preposição
18	numas	Artigo
19	que	Conjunção
20	quando	Conjunção
21	mas	Conjunção
...		

Tabela de Pontuação

Id	Pontuação	Significado	Uniformização
1	...	Reticências	#...#
2	!	Ponto Exclamação	#!#
3	!!	Ponto Exclamação	#!!#
4	!!!	Ponto Exclamação	#!!!#
5	?	Ponto Interrogação	#?#
6	??	Ponto Interrogação	#??#
7	???	Ponto Interrogação	#???#

Tabela de Sinónimos

Id	*Concept*	Sinónimo	Uniformização
1	1kg	1kg	1kg
2	1k	1kg	1kg
3	2kg	2kg	2kg
4	2 k	2kg	2kg
5	caixa	caixa	Caixa
6	caixas	caixa	Caixa
7	msg	mensagem	Mensagem
8	mensagem	mensagem	Mensagem
9	kero	querer	querer
10	sff	se faz favor	se_faz_favor
11	encomendas	encomenda	encomenda
...			

[48] Um *outlier* é um dado significativamente diferente dos restantes dados (Aggarwal, 2017).

[49] *Bag-of-words* é o processo de representar um texto através de um conjunto de conceitos e os mesmos podem aparecer mais de uma vez (Moreira et al., 2019; Provost e Fawcett, 2013; Russell, 2014).

5.3.1.1 Tabela de *smiles*

No conteúdo das mensagens produzidas em linguagem natural encontra-se uma mistura de texto, *smiles*, etc. Entre outros, os utilizadores das RSO utilizam acrónimos, abreviaturas, imagens para comunicar de forma mais rápida. A produção de nexo da comunicação, construída por esta linguagem, consegue-se através da articulação e inserção de alguns (ou todos) destes recursos, no discurso *web*. Os *smiles*, também designados de *emoticons*, fazem parte da linguagem utilizada nas RSO e são incluídos no meio dos textos para expressarem algumas emoções que o texto, por si só, não transmite (pelo menos de forma rápida, sem recurso a adjetivação abundante e figuras de estilo mais elaboradas).

Como as trocas discursivas produzidas no discurso *web* incorporam estes recursos linguísticos, considerou-se importante inserir no modelo uma tabela para os interpretar. A tabela de *smiles* permite interpretar a chamada gíria *online* e detetar sentimentos, positivos ou negativos, do autor da mensagem, relativamente a uma determinada situação ou acontecimento. A tabela é alimentada com recurso a listas de *smiles* do *Facebook* encontradas na *internet*[50]. Por exemplo, *smiles* tais como "**:-)**", "**:)**", "**:]**" e "**=)**" exprimem um sentimento positivo de alegria, enquanto que, em sentido oposto, *smiles* tais como "**:-(**", "**:(**", "**:[**" e "**=(**", expressam um sentimento de tristeza.

5.3.1.2 Tabela de pontuação

A tabela de pontuação é alimentada manualmente, para definir a pontuação que não deve ser eliminada. Por norma a maioria das investigações retiram tudo o que é pontuação. No entanto, considerou-se no modelo desenvolvido que a sequência de alguns caracteres tais como "!!!", "???" e/ou "…", não devem ser retirados pois os utilizadores desenvolveram novas formas de comunicar utilizando a pontuação. O objetivo da utilização destas sequências de caracteres de pontuação é enfatizar uma ideia na mensagem e a sua inclusão na análise permite perceber o grau de satisfação e/ou insatisfação. A utilização de vários sinais de pontuação, como por exemplo "!!??", serve para reforçar indignação e são cada vez mais utilizados no discurso *web*. A incorporação no modelo da tabela de pontuação permite resolver um dos maiores desafios das abordagens que processam dados textuais, que é

[50] Acesso em 25 de junho de 2020, disponível em portais com listas de *smiles*: http://www.symbols-n-emoticons.com/p/facebook-emoticons-list.html e https://www.superdicas.net/internet/significados-dos-smiles.html.

converter um pedaço de texto num vetor e conseguir evidenciar as suas características mais relevantes e importantes.

5.3.1.3 Tabela de *stopwords* e *noise-words*

No modelo, a tabela de *stopwords* e *noise-words*[51] (para palavras irrelevantes) é utilizada no PLN com o objetivo de eliminar as palavras (muito frequentes) que contêm pouca ou nenhuma informação, ajudando a diferenciar o texto onde ocorrem. Uma *stopword* é um conceito (palavra) considerado irrelevante, como por exemplo "a" ou "alguns" que, por serem tão comuns, considera-se que têm valor semântico igual a zero. A lista de *stopwords* do modelo é composta por uma combinação básica de letras e números, bem como pronomes, advérbios, preposições, alguns verbos, adjetivos e conjunções.

Podemos afirmar que a lista de *stopwords* é o universo dos conceitos irrelevantes para cada contexto organizacional e as *noise-words* são a amostra de conceitos irrelevantes existentes em cada mensagem. O motivo para a sua remoção prende-se com o facto de tornarem o texto mais simples, logo mais fácil de analisar. Os conceitos mais comuns a retirar são: preposições (porque exprimem relações entre duas partes de uma frase), conjunções (porque estabelecem a ligação entre dois ou mais elementos e não desempenham funções sintáticas na frase à qual pertencem), artigos definidos (porque têm como objetivo individualizar ou destacar algo) e artigos indefinidos (porque têm como objetivo determinar de forma vaga algo), entre outros, que não acrescentam significado ao texto.

A decisão de quais os conceitos a eliminar depende de cada contexto de aplicação. Por exemplo, de acordo com as listas da *internet* de *stopwords* para o idioma inglês o conceito "*us*" (tradução do pronome pessoal nós) é considerado como sendo um pronome, logo é para eliminar. No entanto, alguns utilizadores utilizam este conceito como acrónimo para se referirem aos Estados Unidos da América. A decisão de eliminar este conceito depende do contexto em análise e que utilização fazemos dos dados.

A tabela é alimentada de duas formas. Uma com os conceitos dos idiomas encontrados nas trocas discursivas e com recurso à literatura e à *internet*. Na *web* facilmente

[51] As *stopwords e noise-words* são conceitos da linguagem natural que podem ser descartados por não terem grande significado para o processamento semântico. Os conceitos que não acrescentam significado ao texto são preposições, conjunções, artigos definidos e indefinidos, números, sinais de pontuação, letras simples (Celko, 1999; Croft et al., 2015; Feldman e Sanger, 2007; Ingersoll et al., 2013).

se encontram listas de *stopword*[52], em vários idiomas, que podem ser incorporadas na *cleaning database*. As tabelas foram ainda alimentadas com os conceitos do contexto em análise, encontrados nas mensagens e identificados com recurso a técnicas de *text-mining*.

5.3.1.4 Tabela de sinónimos

Nesta secção, descreve-se a tabela de sinónimos construída e as técnicas utilizadas para a padronização dos dados do discurso *web*. A tabela de sinónimos tem dois objetivos principais: corrigir erros e uniformizar conceitos com o mesmo significado, quer sejam termos sinónimos quer sejam acrónimos.

O primeiro objetivo é corrigir os erros que possam existir no conteúdo textual das mensagens, quer ortográficos, quer de codificação de fontes (*character sets*). Uma das formas de identificar os erros é através das métricas da SNA. Nos dados textuais os erros têm pouca ocorrência (são *outliers*) e os conceitos corretos têm mais ocorrência. Após calcular as métricas da SNA, se existirem conceitos com *out-degree* = 1 os mesmos devem ser analisados para identificar se estamos ou não perante um erro nos dados. Caso afirmativo a tabela deve ser atualizada para futuramente o descartar e/ou corrigir.

Alguns erros ortográficos são comuns e podem por isso ser incluídos na tabela de sinónimos. Por exemplo, o utilizador digitou *Favebook* em vez de *Facebook*. Este é um erro ortográfico comum, em que o utilizador digitou o caracter "v" que no teclado[53] está do lado direito do "c".

No que se refere à codificação de fontes (*character sets*), como as RSO são alimentadas por utilizadores de várias nacionalidades, existem sempre caracteres "estranhos" nos conteúdos das mensagens. Sempre que possível, esses caracteres "estranhos", quando identificados, devem ser inseridos na tabela de sinónimos para que o algoritmo converta os caracteres ilegíveis no seu equivalente legível. De referir que é esperado que haja conceitos, nos dados textuais, para as quais não há sinónimos.

Para as máquinas, mundo dos *bits* e *bytes*, cada caracter tem um código associado que é interpretado de forma diferente. Para processar de forma automática os dados textuais,

[52] Exemplos: http://dev.mysql.com/doc/refman/5.7/en/fulltext-stopwords.html, http://solariz.de/649/deutsche-stopwords.htm, http://www.ranks.nl/stopwords/portugese, http://www.webconfs.com/stop-words.php, https://www.docear.org/2012/09/28/list-of-6513-stop-words-for-17-languages-english-german-french-italian-and-many-others/. Último acesso em 7 de abril 2020.

[53] Para o alfabeto latino, atualmente o mais utilizado em computadores é o *layout* de teclado QWERTY.

utilizando a *cleaning database*, o *software* tem de interpretar corretamente os caracteres das mensagens (textos). Na Tabela 4, exemplifica-se a interpretação computacional e como são codificados alguns caracteres no formato *1252: Western European (Windows)* e no formato *65001: Unicode (UTF-8)*[54].

Tabela 4 - Exemplo de formatos de *character sets* (conjunto de caracteres)

1252: *Western European* (Windows)	**65001: *Unicode* (*UTF-8*)**
VW_s	VW's
doesn_t	doesn't
World PremiÃ¨re	World Première
Volkswagen dÃ©buts	Volkswagen débuts
was expecting ðŸ˜	I was expecting 😐
Championship Manufacturersâ€™ crown as SÃ©bastien	Championship Manufacturers' crown as Sébastien
werenâ€™t	weren't
Like my Â£35000 Passat	Like my £35000 Passat

Incorporar no modelo uma tabela com sinónimos tem como objetivo utilizar a técnica de *stemming*[55], para agrupar palavras com a mesma origem semântica e aumentar a probabilidade de correspondência, quando utilizadas pelo algoritmo. Quando dois sinónimos são definidos por duas dimensões diferentes, são identificados como dois conceitos distintos e, para além de criarem uma redundância, não há correspondência semântica entre eles. Utilizar a técnica de *stemming* permite juntar formas diferentes da mesma palavra num único recurso textual. Esta técnica foi considerada, porque os utilizadores recorrem a diversas designações para expressarem um mesmo conceito, sendo necessário uniformizar várias ocorrências. Por exemplo, os conceitos *call center*, *callcenter*, *call-centre* e *contact center* uniformizam-se para a designação *call-center*.

Existem, ainda, acrónimos (utilizados na gíria da *web* social) que também foram incluídos na tabela para auxiliar o processamento deste tipo de dados. Por exemplo, os acrónimos conceitos "pq" e "tb" utilizados para designar o conceito "porque" e "também" respetivamente.

A tabela de sinónimos alimenta-se com recurso à identificação dos conceitos que são sinónimos[56] uns dos outros, dentro do contexto de análise. Para a lista de conceitos desta

[54] O *UTF*-8 é um padrão universal para converter caracteres de textos e o *Unicode* é um conjunto de caracteres único criado para interpretar vários idiomas (Croft et al., 2015).

[55] A técnica de *stemming* permite agrupar palavras que têm a mesma origem semântica (Ingersoll et al., 2013).

[56] Um sinónimo é um conceito ou frase que, dentro de um contexto, é substituível por outro conceito ou frase (Croft et al., 2015).

tabela, a adição ou não de uma palavra ou pares de palavras depende em muito da sua utilização no contexto. Alguns conceitos só têm significado num contexto específico ou para um determinado idioma. A diferença de significados entre conceitos dificulta a aplicação de padrões em dados de contextos diferentes. Isto porque a mesma frase ou conjuntos de palavras podem ter significados diferentes dependendo do domínio onde são utilizados.

Os dois principais problemas com conceitos são a equivalência de palavras diferentes (sinónimos) e a ambiguidade das palavras homónimas que representam termos diferentes. Tal implica que a transformação e codificação dos dados textuais seja muitas vezes executada manualmente ou de forma semiautomática.

5.3.2 Algoritmo de transformação semântica

No modelo, a identificação e a uniformização dos conceitos para construção de redes semânticas executam-se com recurso a um algoritmo criado, para o efeito, em *Excel* VBA. No discurso *web*, a semântica encontra-se escondida no seu interior e aqui reside o desafio de a conseguir capturar para que seja utilizada pelas máquinas. Uma das formas de o fazer é através da utilização de métodos formais que transformam a informação implícita (oculto no discurso) em explícita, tornando-a processável pelos computadores.

Para que as máquinas estruturem e interpretem a semântica contida em dados textuais é necessário definir e codificar regras em algoritmos. Ou seja, é preciso ensinar o computador a ler, interpretar e entender a estrutura de uma frase e/ou de um parágrafo. No modelo proposto neste trabalho, a abordagem utilizada para o efeito é relativamente simples.

O algoritmo para transformação semântica foi construído em cima de um conjunto de técnicas, onde se incluem a teoria da informação, PLN, *data-mining* e *text-mining*. Dentro do âmbito dos métodos de análise de dados, o *data-mining* e o *text-mining* podem ser considerados abordagens exploratórias de descoberta de conhecimento. Assim, no modelo, o algoritmo permite filtrar e refinar os dados textuais para torná-los utilizáveis e pesquisáveis. Isto é, transformam dados não estruturados em dados estruturados.

Após a limpeza dos dados textuais é possível criar redes semânticas, com resumos das trocas discursivas e redes de conceitos, que se encaixam numa estrutura analítica mais ampla de apoio à decisão. O algoritmo desenvolvido em *Excel VBA* é uma macro que executa tarefas específicas, automatizadas e, quando ativada, executa uma sequência de instruções. O algoritmo criado para processamento e extração de conceitos envolve os seguintes passos:

1.º *Input* e *Output*: Conjunto de instruções do algoritmo que permite selecionar a área de entrada (*input*) onde estão as mensagens e a área de saída (*output*) onde os conteúdos das mensagens processadas serão armazenados. Neste passo, o algoritmo lê e transforma o conteúdo de cada mensagem em tantos pedaços quantos os conceitos existentes, como exemplificado na Figura 33. Após seleção da área de *input* e da área de destino dos conceitos processados, o algoritmo executa as rotinas de limpeza de dados e insere um cabeçalho com os atributos dos mesmos. Nas linhas abaixo do cabeçalho são inseridos os conceitos extraídos das mensagens;

2.º O processo de limpeza de dados (*data cleaning*) desencadeia-se através de uma rotina designada de *DecodePost* que limpa e uniformiza os conceitos invocando várias outras rotinas com utilização da *cleaning dataBase*:

a) A sub-rotina *RemoveSymbols* é composta por um ciclo que retira qualquer caracter que não seja alfanumérico ou pontuação. O ciclo, através de um contador, permite repetir um conjunto de instruções tantas vezes quanto a quantidade de caracteres existentes em cada mensagem. Para validar os caracteres a eliminar, definidos como critério no ciclo, os mesmos são comparados com os da tabela *American Standard Code for Information Interchange* (ASCII)[57]. Por exemplo, se existirem nas mensagens os caracteres "[", "\", "]" e "^" com valor 91, 92, 93 e 94 respetivamente na tabela ASCII, os mesmos são eliminados;

b) A sub-rotina *ReplaceSmiles* dimensiona um *array*[58] que guarda os *smiles*, conjunto de caracteres definidos na tabela de *smiles* da *cleaning database*, para os encontrar na mensagem. Isto é, utilizando um ciclo, procura os *smiles*, compara-os com os termos definidos e substitui-os pelo termo de uniformização. O ciclo, através de um contador, repete um conjunto de instruções até não encontrar mais nenhum conjunto de caracteres que seja igual aos *smiles* definidos. Por exemplo, os conjuntos de caracteres ":-)" e ":-(" são substituídos pelo termo definido como padrão "(*Smiley*)";

[57] ASCII é um padrão de codificação de caracteres utilizado para validar texto (Croft et al., 2015).

[58] Um *array* é um grupo de elementos do mesmo tipo que tem um nome comum (Alexander e Kusleika, 2016).

c) A sub-rotina *ReplacePontuacao*, através de um ciclo, procura em cada mensagem a pontuação existente na *cleaning database* e substitui-a pelo termo de uniformização. Por exemplo, os conjuntos de caracteres ".", "?" e "!" são substituídos pelo termo padronizado "#.#", "#!#" e "#?#" respetivamente. Quando o número de caracteres de pontuação é superior a dois, o algoritmo está programado para assumir sempre três caracteres de pontuação. Esta sub-rotina valida a quantidade sequencial de caracteres de pontuação que encontra, para padronizá-los de acordo com os casos definidos. Por exemplo, se o ciclo encontrar três, quatro ou mais caracteres seguidos com um ponto de exclamação, padroniza para "#!!!#". A pontuação normalmente marca o limite das frases, mas, no que diz respeito ao discurso da *web*, pode ser utilizada para enfatizar o que está a ser escrito e, portanto, é uma exceção no uso da linguagem;

d) A sub-rotina *RemoveStopword* é de grande importância, pois no processo de limpeza de texto há conceitos utilizados com frequência que devem ser filtrados. Por exemplo, conceitos tais como "a", "de", "a", entre outros. Como já referido estes conceitos são designados de *stopword* e *noise-word* e têm definições e significados diferentes sendo importante fazer a sua distinção. A sub-rotina utiliza a lista de conceitos padronizados como *stopword*, para os encontrar na mensagem e eliminar. Por exemplo, na frase "Será que me podes arranjar 2kg de morangos para hoje? Obrigada", os conceitos "que", "me", "de", "para" e "obrigada" podem ser removidos. A eliminação destes conceitos não retira o sentido da ideia contida na mensagem. O resultado final da frase ficará: "podes arranjar 2kg morangos";

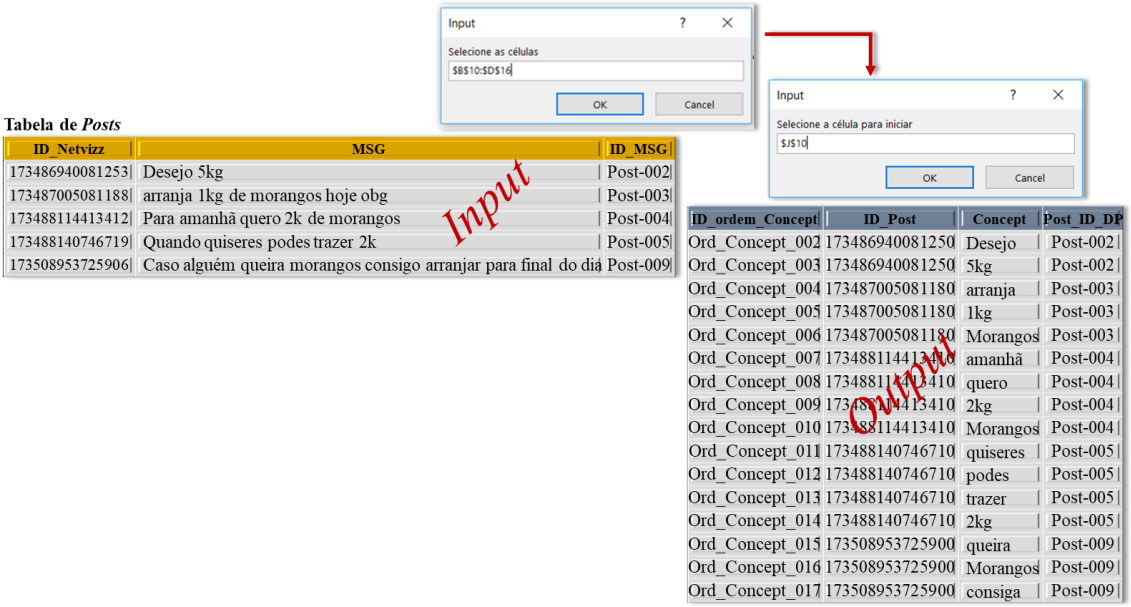
e) A sub-rotina *ReplaceSinonim* é também de extrema importância, pois utiliza-se para executar de forma automática algumas das tarefas de *semantic parsing*[59], como a normalização de conceitos sinónimos entre si, tais como nomes, endereços, medidas, empresas, datas, etc. Por meio de um ciclo, verifica se a mensagem tem conceitos que existem na lista de termos a substituir e, em caso afirmativo, normaliza-os. Por exemplo, os conceitos "Volkwagen", "Volkswagen's", "VW", "VWs", e "#Volkswagen" são

[59] *Semantic parsing* é uma técnica do PLN para normalização de conceitos sinónimos (Ingersoll et al., 2013).

normalizados para "Volkswagen". Esta normalização reduz o volume de dados semânticos obtidos;

3.º Por último, no final de todo o processo de transformação semântica, o algoritmo estrutura os dados em formato tabular. As técnicas de PLN utilizadas pelo algoritmo para estruturar as mensagens incluem *tokenization*[60] e as sequências *N-gram*[61]. A *tokenization* tem como objetivo "partir" o texto em tantos pedaços quantos os existentes nas mensagens, de acordo com sequências *N-gram*. Neste modelo essas sequências são *unigrams*: palavras individuais separadas pelo algoritmo.

Figura 33 - *Input e output* de dados



5.3.3 Output do processamento semântico

O *output* final, obtido após execução de todas as etapas da terceira fase, é um conjunto de dados estruturados que, por um lado, alimentam a *cleaning database* com conceitos padronizados e, por outro, servem de *input* para a *graph database*. Ou seja, os *outputs* obtidos no final do processo têm duas finalidades: (i) a criação/obtenção de padrões para o processamento semântico; e (ii) a criação/obtenção de dados estruturados e

[60] A *tokenization* envolve "partir" frases, parágrafos ou seções de um documento em múltiplos pedaços também designados de *token* (Ingersoll et al., 2013).

[61] *N-gram* refere-se à sequência de n itens formados por sílabas, letras, conceitos ou estruturas mais complexas (Provost e Fawcett, 2013).

padronizados para criar redes semânticas. Nas áreas de PLN e de recuperação de informação, estes conjuntos de dados textuais são designados por *bag-of-words*. A abordagem *bag-of-words* é relativamente simples e não requer uma análise linguística muito sofisticada, por esse motivo foi a opção escolhida como uma das técnicas de *text-mining*.

Do lado da *cleaning database*, utilizar *bag-of-words* permite lidar com os aspetos semânticos para o processamento do texto contido nas mensagens. Na *cleaning database*, as *bag-of-words* são apenas um conjunto de conceitos individuais onde a gramática, a ordem das palavras e a estrutura das frases são ignoradas. Cada conceito da *cleaning database* funciona como uma palavra-chave, que entra no algoritmo como parâmetro para o processamento semântico.

Relativamente às *bag-of-words* que alimentam a *graph database*, os conceitos nelas contidos servem para criar resumos do discurso *web* e identificar palavras-chave. Para criar os resumos, como referido anteriormente, o algoritmo atribui *IDs* sequenciais aos conceitos. São criados tantos *IDs* quantas as ocorrências do conceito na mensagem, ou conjunto de mensagens. Se o conceito existir 10 vezes, teremos 10 *IDs* distintos. Assim, recorrendo na fase de estruturação a técnicas de *data-mining* é possível reconstruir os resumos das trocas discursivas. No caso da identificação de palavras-chave é atribuído um *ID* único a cada conceito. Ou seja, se o conceito existir 10 vezes, ser-lhe-á atribuído um *ID* único.

5.4 Resumo

As fases do modelo proposto neste trabalho de investigação foram descritas neste capítulo. O modelo apresenta uma estrutura para extrair, armazenar, processar, estruturar e analisar dados de RSO de um contexto organizacional específico. A sua estrutura está dividida em três fases: (i) Extração de Dados; (ii) Processamento de Dados; e (iii) Interpretação e Processamento Semântico. A implementação do modelo faz-se com recurso a *software* de utilização comum, técnicas da SNA para visualização da rede e cálculo de métricas e técnicas de PLN tais como *data-mining*, *text-mining*, *data cleaning*, entre outras.

Os dados das três entidades do discurso *web* (*user*, *post* e *concept*) são encadeados e utilizados para a criação de redes, que permitem ter uma visão global não só das interações sociais entre utilizadores, mas também dos diálogos produzidos.

O modelo fornece um método sistemático que permite criar e analisar várias redes, quer ao nível dos utilizadores quer ao nível da semântica. A análise dos dados semânticos

tem por finalidade a criação de um *input* de dados para alimentar a *cleaning database* (com recurso à técnica *unsupervised* de extração de conceitos dos dados não estruturados do conteúdo das mensagens) e responder às questões e aos problemas de decisão de um contexto organizacional específico (utilizando-se dados padronizados e uniformizados para suporte à extração de informação útil para apoio à decisão).

O modelo integra ainda uma *graph database* que organiza os dados em formato grafo para criação e visualização de possíveis sub-redes, de acordo com o objetivo da análise.

6 Estudos de caso

A metodologia de investigação *action design research*, referida no subcapítulo 1.4, articula os benefícios não só da *design science research*, mas também da *action research* para projetar um artefacto, fundamentado na teoria académica que resolve uma questão prática (Mettler, 2015; Mullarkey e Hevner, 2018). De acordo com a literatura, nesta metodologia, a avaliação é definida como um processo exigente de teste e verificação do comportamento do artefacto no ambiente para o qual foi idealizado e projetado.

Para avaliar o artefacto, o modelo proposto no capítulo anterior, e a solução que o mesmo oferecia para a resolução do(s) problema(s) prático(s) em situações reais de apoio à decisão, realizaram-se quatro estudos de caso. Estes permitiram o aperfeiçoamento do modelo e a sua aplicabilidade em diferentes cenários, com níveis de participação distintos. Foram ainda utilizados para perceber como estruturar, de forma simples, dados das RSO, utilizando a metodologia SNA, de modo a extrair informação do discurso *web* produzido pelos seus utilizadores para o apoio à decisão.

Os quatro estudos de caso apresentam-se na seguinte sequência: (1) estudo de caso "Leitões" (Antunes et al., 2018; Freire et al., 2015b); (2) estudo de caso "TAP" (Antunes et al., 2018; Freire et al., 2015c); (3) estudo de caso "Morangos" (Antunes et al., 2018; Freire et al., 2017, 2021) e (4) estudo de caso "Volkswagen".

Apesar da aplicabilidade da estrutura do modelo não depender de uma RSO específica, utilizou-se o *Facebook* para todos os estudos de caso apresentados. O *Facebook* é uma RSO com alguma maturidade, robusta e bem documentada, relativamente à forma como as trocas discursivas se encontram estruturadas (Russel, 2019). Para a extração de dados utilizou-se o *Netvizz* e o *NodeXL* recolhendo-os de duas *FanPages* e de dois grupos criados e geridos pela autora. O *software Gephi* foi utilizado para visualizar e manipular as redes, bem como para o cálculo das métricas mais importantes da SNA.

De referir que, nos grupos criados, conheciam-se os utilizadores, o contexto de decisão e a linguagem utilizada. Nas *FanPages* da Companhia Aérea TAP e do fabricante de veículos *Volkswagen* desconhecia-se, à partida, que tipo de dados se iriam encontrar e que informação iriam "mostrar". Na Tabela 5, evidenciam-se as principais características dos quatro estudos de caso.

Tabela 5 - Estudos de caso: características

#	Estudo de caso	*Software* de recolha de dados	Tipo de RSO	Contexto organizacional	Tipo de decisão
1	Leitões	*NodeXL* *Netvizz*	Grupo	Conhecido	Dados direcionados para decisão definida *a priori*: logística evento.
2	TAP	*NodeXL*	*FanPage*	Desconhecido	Dados exploratórios para perceber o que diziam os clientes e se era importante atuar e em que sentido.
3	Morangos	*Netvizz*	Grupo	Conhecido	Dados direcionados para decisão definida *a priori*: estimar quantidades produto.
4	VW	*Netvizz*	*FanPage*	Desconhecido	Dados exploratórios para monitorizar "conversas" *online*.

Como a interpretação da categorização pequeno ou grande pode induzir incerteza, no que se refere à dimensão dos dados dos estudos de caso, na Tabela 6 apresenta-se uma caracterização quantitativa. Esta classificação caracteriza as amostras de dados dos estudos de caso de acordo com duas perspetivas: a das ciências sociais e a das ciências da computação. A primeira, quantifica de acordo com a dimensão das redes (quantidade de nós e de ligações). A segunda, caracteriza o tamanho das amostras de dados sociais de acordo com a unidade de medição da informação ou de armazenamento (*megabyte*) dos sistemas computacionais. Em ambos os casos quantificam-se os dados no momento da recolha (originais) e após o seu processamento (finais).

Tabela 6 - Caracterização da dimensão dos estudos de caso

			Perspetiva das ciências sociais				Perspetiva das ciências computacionais	
Estudo de caso	Período temporal		Nós		Ligações		*Megabyte*	
	Início	Fim	Originais	Finais	Originais	Finais	Originais*	Finais*
Leitões	22/nov/14	11/jan/15	107	246	314	469	0,056 Mb	1,251 Mb
TAP	01/mai/15	12/mai/15	2.205	26.510	10.591	51.614	4,749 Mb	58,481 Mb
Morangos	10/mai/16	16/mai/16	59	368	119	693	0,053 Mb	9,374 Mb
VW	31/ago/15	31/jan/16	32.486	196.768	36.003	356.438	12,072 Mb	1.196,316 Mb

* Fonte de dados: GDF, CSV, XLS

A análise quantitativa das métricas de centralidade das RSO, pode envolver diferentes escalas de observação, que vão desde a estrutura global da rede à posições individual de um nó, podendo incluir subgrupos (Hanneman e Riddle, 2005). Assim, as redes analisadas não tiveram apenas uma escala de análise, mas várias, porque tanto a estrutura da rede como a posição de um nó podem caracterizar-se através de várias métricas. Para as métricas *degree centrality*, *in-degree centrality* e *out-degree centrality*, em que os resultados da SNA eram valores absolutos, a escala podia variar de zero ao número máximo de ligações

que um nó tinha dentro da rede. Já para as métricas *closeness centrality*, *betweenness centrality*, *eigenvector* e *PageRank*, como os resultados foram normalizados, a escala variava entre zero e um.

6.1 Estudo de caso – Leitões

O estudo de caso designado de "Leitões" teve como objetivo principal perceber se era possível estruturar, de forma simples, as trocas discursivas da rede social *Facebook*, utilizando a SNA, para extração de informação para apoio à decisão. Este estudo de caso foi o primeiro, tendo sido realizado para caracterizar e identificar as áreas de conhecimento e os recursos necessários para analisar os dados provenientes das RSO para apoio à decisão. O estudo de caso, abordou a complexidade subjacente à análise do discurso *web*, focou os aspetos importantes e a considerar na sua análise, bem como a sua utilização no apoio à decisão no contexto das RSO.

As RSO caracterizam-se por serem muito dinâmicas, crescerem e alterarem-se à medida que surgem novos nós, fruto quer de novas interações sociais, quer da publicação de novas mensagens. Por esse motivo considerou-se importante delimitar o objeto de estudo apresentado neste subcapítulo criando-se um grupo fechado na RSO *Facebook*. A rede criada pela discussão do tema foi analisada a partir da perspetiva *ego-centric*, onde os utilizadores estavam definidos pelas relações específicas que tinham com o *ego*[62].

A criação do grupo teve ainda como finalidade a utilização de uma rede de pequenas dimensões e a recolha de dados reais. Utilizar uma rede pequena permitiu limitar a quantidade de elementos de cada entidade do discurso e, assim, não depender da complexidade e limitações associadas a grandes volumes de dados. Para além disso, o facto de a autora fazer parte do grupo, garantiu que a linguagem utilizada e a semântica associada eram conhecidas. Por esta razão, não foram necessárias ontologias para interpretar os conceitos contidos nos *posts* e enquadrar e decifrar o que estava implícito no discurso.

A abordagem *ego-centric* mostrou-se particularmente útil para limitar quer o objeto de estudo, quer as fronteiras da rede analisada. A rede *ego-centric*, que oferece uma visão da rede a partir da perspetiva de um único nó (*ego*) e das suas ligações, permitiu que fosse possível uma abstração (em parte) de alguma da complexidade associada à análise do

[62] No estudo de caso dos Leitões o *ego* definido foi a autora.

discurso *web* referida no capítulo 2, justificada pelo facto da autora, enquanto analista, fazer parte do contexto de decisão. Conheciam-se os intervenientes, o tema em discussão, o contexto, a semântica e a linguagem utilizada, uma vez que o grupo era constituído por utilizadores conhecidos.

Este grupo era composto por colaboradores de um departamento de uma empresa nacional portuguesa, dispersa geograficamente por todo o país. Os colaboradores, a convite da autora, aderiram voluntariamente ao grupo, evidenciando diferentes níveis de participação. Assim, foi possível estudar a rede ao nível da sua composição, da sua estrutura e das suas características (através de métricas aplicadas a cada entidade do discurso). Na sequência de uma reunião anual da empresa, colocou-se em discussão no grupo a organização do almoço. À organização de um simples almoço está associado um processo logístico. Quem organiza precisa de saber quem estará presente, bem como as preferências para o tipo de comida e espaço de forma a satisfazer os participantes. Qualquer que seja a problemática, uma tomada de decisão racional exige uma pesquisa das alternativas sugeridas e das informações dadas sobre as preferências, para obtenção dos resultados pretendidos.

Neste estudo de caso, analisaram-se os dados recolhidos considerando quatro perspetivas diferentes: (1) a rede com todos os utilizadores que compunham o grupo; (2) a rede com os utilizadores que tiveram interação com outro elemento do grupo; (3) a rede com todos os utilizadores e respetivos *posts*; e (4) a rede semântica.

Os dados analisados, fruto das contribuições publicadas pelos utilizadores, foram recolhidos entre 22 de novembro de 2014 e 11 de janeiro de 2015. Utilizou-se o *Gephi* como ferramenta de SNA, evitando a utilização de uma linguagem de programação para manipulação e visualização de dados e simplificando a análise das quatro redes. Utilizando essa ferramenta, para as redes 1 e 2 calcularam-se as métricas de centralidade da SNA nomeadamente o *degree centrality*, *in-degree centrality*, *out-degree centrality*, *closeness centrality*, *betweenness centrality*, *eigenvector* e *PageRank*. As redes 3 e 4 caracterizavam-se por serem redes *two-mode*, compostas por mais do que um tipo de entidade. Isto é, a rede 3 continha três tipos de nós: *users*, *posts* com fotos e *posts* só com texto. No caso da rede 4, esta era composta por dois tipos de nós: *posts* e conceitos. Assim, para estas duas últimas redes, calcularam-se as métricas *in-degree centrality*, *out-degree centrality* e quantidade de *likes* recebidos. Para todas as redes (1, 2, 3 e 4) calculou-se a métrica *modularity class* que permitiu identificar e evidenciar subgrupos, representando-os por cores. Cada uma das redes é composta por um número diferente de subgrupos (4, 2, 6 e 8, respetivamente)

6.1.1 Extração de dados

A extração de dados foi inicialmente executada com recurso ao *NodeXL* por ser uma ferramenta sem custos e que de forma simples e rápida permitiu importar os dados do *Facebook*. Todavia, essa ferramenta demonstrou limitações associadas quer ao conjunto de dados, quer ao tipo de redes que se pretendiam extrair. Para além disso, o *output* de dados devolvido pela ferramenta não permitia o processamento semântico do conteúdo dos *posts* (dados não estruturados). Os dados foram também extraídos do *Facebook* através do *Netvizz*. Comparativamente com o *NodeXL*, verificou-se que o *Netvizz* respondia melhor às necessidades e aos objetivos deste trabalho de investigação.

A ferramenta de extração de dados *Netvizz* permitiu recolher três conjuntos distintos de dados, já estruturados em formato grafo. Estes dados originaram a criação de três redes também elas distintas:

- Rede 1: Continha todos os utilizadores que pertenciam ao grupo, independentemente de existir ou não interação entre eles. Nesta rede, o facto de um qualquer utilizador pertencer ao grupo significou que existia uma relação entre utilizadores. Logo, interpretou-se essa relação entre dois quaisquer utilizadores como sendo uma ligação entre eles;
- Rede 2: Continha todos os utilizadores com interação social. A rede só continha os pares de utilizadores que tinham estabelecido trocas discursivas uns com os outros. Esta rede era constituída por 16 nós e 32 ligações entre eles;
- Rede 3: Continha todos os utilizadores com interação social e as respetivas mensagens e comentários (*posts*). O conjunto de dados desta rede, incluía um ficheiro em formato TSV que continha os conteúdos das mensagens. Esta rede era composta por 64 nós e 142 ligações entre eles.

6.1.2 Processamento e interpretação dos dados

Os dados recolhidos via *Netvizz* foram utilizados como *input* para o *Gephi*, no sentido de calcular métricas da SNA e de visualizar as redes. Os três conjuntos de dados recolhidos já se encontravam em formato g*raph database*, isto é, cada um era composto por duas tabelas: uma para a identificação dos nós e a outra com as ligações entre eles.

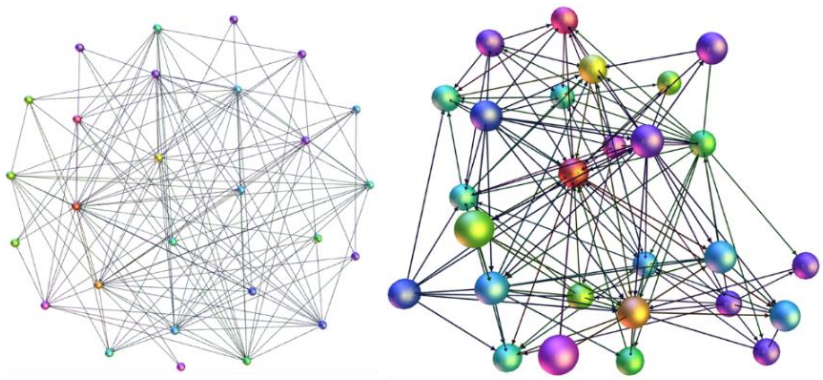
Como o estudo de caso pretendia ir para além da análise da posição estrutural dos utilizadores, extraiu-se o texto dos conteúdos das trocas discursivas e criou-se uma rede semântica. Essa rede semântica (rede 4) construiu-se agregando os dados dos *posts* da rede 3 e os dados não estruturados extraídos desses *posts*. O objetivo foi resumir o texto das trocas discursivas através de palavras-chave.

O *Gephi*, para além de ser um software analítico, possui um módulo de visualização de redes que permitiu atribuir cores, tamanho e outras propriedades aos nós e ligações das redes. Esta funcionalidade permitiu explorar e interpretar os dados através de vários *layouts*. Para além disso, o *Gephi* possibilitou uma manipulação simples das redes e o cálculo das métricas mais importantes da SNA. Exemplo disso são os grafos e as tabelas de resultados apresentados neste estudo de caso e que caracterizam as entidades de cada rede, analisadas a seguir.

6.1.2.1 Rede 1 - Ligações entre utilizadores do grupo

Num primeiro passo, os dados originais da rede 1, sem qualquer processamento, foram diretamente importados para o *Gephi* para serem visualizados (Figura 34 (a)). No sentido de melhorar a representação gráfica dos dados alterou-se o *layout* do gráfico para o *Force Atlas 3D*[63]. O algoritmo deste *layout* aproxima os nós diretamente ligados entre si e afasta os restantes. Esta funcionalidade auxiliou na interpretação da rede e, após a sua aplicação, o grafo convergiu para o estado mostrado na Figura 34 (b).

Figura 34 - Rede 1: representação gráfica dos dados não processados



(a) Dados não processados (b) Dados após aplicação do *layout Force Atlas 3D*

Pela visualização da figura, a única informação possível de observar foi a densidade da rede, ou seja, a existência de vários nós todos interligados uns aos outros. Após esta

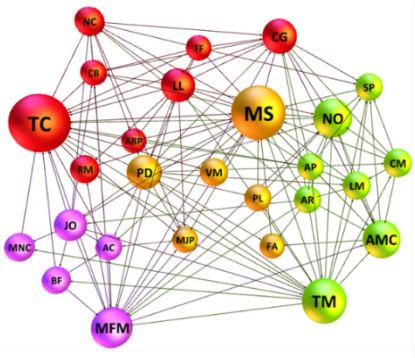
[63] Mais informação sobre o algoritmo *Force Atlas 3D* pode ser consultada em Cherven (2013; 2015).

primeira exploração dos dados com recurso aos métodos de visualização gráfica da SNA, o segundo passo foi o seu processamento. Esse passo, permitiu tornar os nós da rede mais percetíveis, atribuindo-lhes legendas, para facilmente identificar quem estava ligado a quem, se existiam subgrupos, e quem eram os utilizadores mais importantes. Para o efeito calcularam-se as métricas da SNA anteriormente referidas.

Após o cálculo das métricas configurou-se novamente o *layout* do grafo com os resultados obtidos como ilustrado em (a) na Figura 35. O tamanho de cada nó dimensionou-se de acordo com a métrica *betweenness centrality* e a métrica *modularity class* foi utilizada para codificar a cor dos quatro subgrupos existentes na rede. Na Figura 35 mostram-se todos os elementos do grupo, que tinham pelo menos uma ligação a outro qualquer elemento, ou seja, 27 nós e 140 ligações. A tabela (b) da Figura 35 mostra os resultados obtidos para as várias métricas no decorrer da análise dessa rede.

De acordo com os resultados obtidos, para esta topologia de rede, observou-se que os utilizadores TM, PD e AP eram os elementos mais ativos, porque tinham um valor de *out-degree* superior aos outros (14 ligações). Isso indicou que criaram vínculos entre 14 outros utilizadores. Em sentido oposto, verificou-se que dois dos utilizadores (AC e CB) foram pouco ativos pois registaram um valor nulo nessa métrica.

Figura 35 - Representação e resultados da interação entre utilizadores



(a) Rede *one-mode*

Nó	*In-degree*	*Out-degree*	*Betweenness*	*Closeness*	*Eigenvector*	*PageRank*
MS	21	3	0,047	1,000	1,000	0,172
AC	7	0	0,000	0,000	0,754	0,145
CB	8	0	0,000	0,000	0,679	0,083
MJP	7	1	0,001	1,000	0,625	0,079
TC	13	6	0,058	1,000	0,398	0,059
SP	8	1	0,000	0,571	0,361	0,044
NO	10	4	0,023	0,625	0,272	0,038
AMC	8	3	0,020	0,667	0,299	0,035
RM	4	1	0,000	0,571	0,273	0,027
CG	8	6	0,017	0,688	0,154	0,026
LL	7	7	0,015	0,667	0,117	0,025
FA	3	1	0,000	0,571	0,125	0,025
MFM	7	10	0,026	0,813	0,086	0,024
VM	1	1	0,000	0,571	0,151	0,021
LM	5	4	0,001	0,550	0,055	0,019
TM	4	14	0,029	0,792	0,031	0,017
NC	3	8	0,008	0,652	0,022	0,017
ABP	3	3	0,000	0,481	0,028	0,016
MNC	2	1	0,000	0,538	0,034	0,015
PD	2	14	0,018	0,719	0,007	0,015
JO	2	9	0,008	0,611	0,015	0,015
PL	2	2	0,000	0,478	0,030	0,015
AR	1	10	0,003	0,622	0,003	0,014
AP	0	14	0,000	0,694	0,000	0,013
FF	0	8	0,000	0,558	0,000	0,013

(b) Métricas SNA: resultados

Na métrica *in-degree*, o utilizador MS foi o elemento com o valor mais elevado (21 ligações). Em parte, este resultado justifica-se pelo facto de o utilizador MS ser o moderador

do grupo e estarem a ser quantificadas todas as ligações da rede quer tivessem ou não ocorrido uma troca disursiva. Este resultado, significou também que os *posts* de MS foram os mais visualizados ou comentados, porque eventualmente foram considerados mais interessantes e/ou relevantes pelo grupo. O utilizador MS caracterizava-se também por registar valores elevados nas métricas *closeness centrality*, *eigenvector centrality* e *PageRank* (1; 1 e 0,172 respetivamente). O facto deste utilizador possuir um valor elevado de *closeness centrality* representou independência, e que comunicava com os outros elementos do grupo, através de um número mínimo de intermediários. Por definição, os utilizadores que possuem mais ligações tendem a ser mais poderosos, porque podem afetar de forma direta os outros elementos do grupo. Os resultados obtidos através da *eigenvector centrality*, confirmam que MS era um utilizador importante na rede, visto que por definição a métrica quantifica a importância de um nó, em função da importância dos seus vizinhos. Mesmo que o utilizador estivesse ligado só a um outro elemento, tendo assim um valor baixo de *degree centrality*, se esse elemento fosse relevante, por consequência, ele também seria importante.

Relativamente à métrica *betweenness centrality*, verificou-se que o utilizador TC evidenciava o valor mais elevado. Este tipo de utilizador designa-se de utilizador-chave com poder de influência sobre os outros. Este facto foi importante para a análise do discurso *web*, pois este tipo de utilizador tem mais ou menos poder, quando é capaz de influenciar (mais ou menos) os atos e as decisões de outros. A visualização quer do grafo apresentado em (a), quer das métricas da SNA apresentadas em (b), na Figura 35, permitiram de forma rápida e simples caracterizar a importância dos utilizadores acima referidos dentro do grupo.

6.1.2.2 Rede 2 - Utilizadores com interação dentro do grupo

O conjunto de dados da rede 2 também foi importado para o *Gephi* e explorado sem qualquer processamento. Num primeiro passo, representado na alínea (a) da Figura 36, visualizou-se a rede e, para melhorar a sua representação gráfica, alterou-se o *layout*. Para esta rede escolheu-se o *layout* produzido pelo algoritmo *Fruchterman Reingold*[64], que distribui os nós num plano bidimensional e que, apesar de algum nível de separação entre eles, mantém mais próximos os nós diretamente interligados entre si. Após a aplicação do

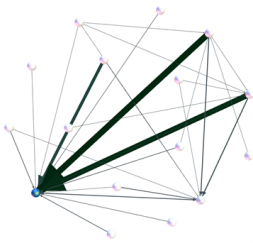
[64] Algoritmo de visualização gráfica apresentado por Fruchterman e Reingold (1991) e incorporado nos *layouts* do *Gephi*.

algoritmo não se verificaram grandes alteração no grafo e avançou-se de imediato para o cálculo das métricas da SNA, apresentadas na Figura 36 em (b), para caracterizar os elementos da rede.

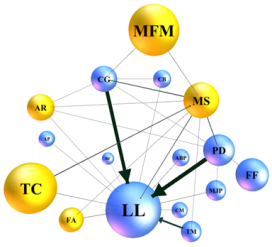
Como ilustra o grafo apresentado em (a) da Figura 36, a única informação que claramente sobressaiu foi a de que todos os nós estavam interligados uns aos outros e não existiam nós soltos. Para além disso, pela dimensão das linhas verificou-se a existência de uma interação forte entre alguns elementos do grupo.

Também nesta rede, o segundo passo foi tornar os nós mais percetíveis para identificar melhor as interações, a existência de subgrupos, e identificar os utilizadores mais importantes. Após o cálculo das métricas da SNA, apresentadas na tabela (b) da Figura 36, reconfigurou-se o *layout* do grafo com os resultados obtidos. Como se pode observar pelo grafo apresentado em (c) na Figura 36, identificaram-se dois subgrupos distintos representados pela cor amarela e azul. O tamanho de cada nó da rede 2 dimensionou-se de acordo com a métrica *betweenness centrality*.

Figura 36 - Representação dos dados da rede 2 pré e pós processamento



Nó	*In-degree*	*Out-degree*	*Betweenness*	*Closeness*	*Eigenvector*	*PageRank*
ABP	0	1	0,000	0,571	0,000	0,009
AP	0	1	0,000	0,571	0,000	0,009
AR	0	4	0,000	1,000	0,000	0,009
BF	0	1	0,000	0,571	0,000	0,009
CB	0	2	0,000	0,667	0,000	0,009
CG	5	2	0,007	0,750	0,435	0,134
CM	0	1	0,000	0,571	0,000	0,009
FA	0	2	0,000	0,667	0,000	0,009
FF	0	1	0,000	0,571	0,000	0,009
LL	13	3	0,117	1,000	0,724	0,298
MFM	0	2	0,000	0,571	0,000	0,009
MJP	0	2	0,000	0,667	0,000	0,009
MS	9	2	0,005	0,600	1,000	0,347
PD	4	3	0,019	1,000	0,307	0,108
TC	0	2	0,000	0,667	0,000	0,009
TM	0	2	0,000	0,667	0,000	0,009



(a) Dados não processados (b) Métricas da SNA (c) Dados processados

A rede 2 representada em (c), continha todos os elementos do grupo que tinham pelo menos uma interação com outro qualquer elemento, ou seja, era constituída por 16 nós e 31 ligações. A diferença entre a rede 1 e esta última é o facto de que a primeira representava todos os 27 elementos do grupo quer tivessem ou não interagido com outros elementos. A segunda rede representava os elementos do grupo que realizaram pelo menos uma troca discursiva com outro qualquer elemento. De acordo com os resultados da tabela (b), apresentados na figura, observa-se que os utilizadores com valores mais elevados, na rede 2, nas diferentes métricas foram MS, LL e AR.

O utilizador MS registou o valor mais elevado nas métricas *eigenvector centrality* e *PageRank* (1 e 0,347 respetivamente). O utilizador LL era o elemento com valor mais elevado na métrica *in-degree*, igual a 13 ligações e na métrica *betweenness centrality* registou um valor de 0,117. Estes resultados denotam poder do utilizador no controlo do fluxo de informação. Pois se por um lado tinha um *in-degree* elevado que evidenciava a importância dos seus *posts*, por outro a métrica *betweenness centrality* indicava que o facto de estar entre a comunicação de outros lhe conferia poder.

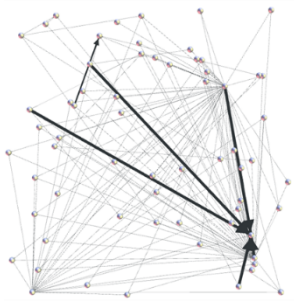
O utilizador AR tinha um *out-degree* igual a 4 ligações, o que indicou que interagia com quatro outros elementos da rede. O facto de AR, LL e PD registarem o valor mais elevado da métrica *closeness centrality*, indicou que comunicavam com muitos elementos, através de um número mínimo de intermediários. Logo, rapidamente faziam chegar informação aos outros elementos. Como referido no capítulo 3, por definição, esta métrica quantifica as ligações de saída de um nó e está relacionada com o tempo que uma informação demora a ser partilhada por todos os nós da rede.

6.1.2.3 Rede 3 - Utilizadores com interação e respetivos conteúdos

Enquanto que as duas redes anteriores se caracterizavam por serem redes *one-mode*, só com utilizadores, a rede 3 caracterizava-se por ser do tipo *two-mode*. Esta rede, era composta por todos os elementos do grupo, que tinham pelo menos uma interação com outro qualquer elemento e com o conteúdo publicado na sequência das interações. A rede era composta por dois tipos de nós: *users* e *posts* com fotos e/ou texto, ou seja, continha 142 ligações entre 64 nós, dos quais 14 eram utilizadores e 50 eram *posts*.

Também nesta rede, os dados originais foram analisados sem qualquer processamento e importados para o *Gephi*. Como ilustra a Figura 37, num primeiro passo interpretou-se a rede utilizando o grafo apresentado em (a). De seguida, avançou-se para o cálculo das métricas da SNA, apresentadas em (b) e por último, para melhorar a representação dos nós da rede, alterou-se o *layout* do grafo para *Fruchterman Reingold*, alínea (c). Após a aplicação das métricas da SNA verificaram-se alterações na disposição e dimensão dos nós do grafo.

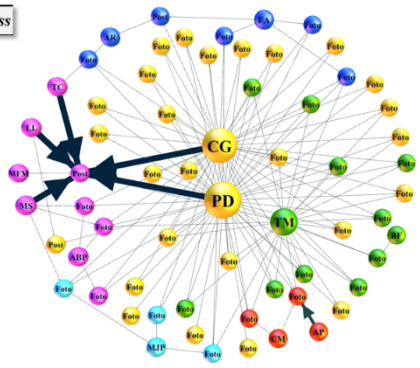
Figura 37 - Representação dos dados da rede 3 pré e pós processamento



(a) Dados não processados

Nó	*In-degree*	*Out-degree*	*likes*	*Modularity Class*
ABP	0	3	3	4
AP	0	1	1	0
AR	0	2	2	5
BF	0	1	1	1
CG	0	44	44	2
CM	0	2	1	0
FA	0	4	4	5
LL	0	2	1	4
MFM	0	1	0	4
MJP	0	3	3	3
MS	0	4	2	4
PD	0	49	48	2
TC	0	2	2	4
TM	0	24	23	1

(b) Métricas da SNA



(c) Dados processados

Pela análise do grafo, apresentado em (c) na figura, observou-se que todos os nós estavam interligados uns aos outros e a dimensão das ligações indicava uma interação mais forte entre alguns nós. Assim, o passo seguinte foi perceber que tipo de nós tinham mais ligações e quais os mais importantes.

Após o cálculo das métricas da SNA, configurou-se novamente o *layout* do grafo com os resultados obtidos. O grafo apresentado em (c) na Figura 37, ilustra os resultados obtidos e permitiu identificar as entidades mais importantes na perspetiva da SNA. Com recurso à métrica *modularity class* identificaram-se seis subgrupos na rede, cada um com uma cor atribuída, conforme representado no grafo. As diferentes cores, para além de caracterizarem os subgrupos, representam os nós mais densamente ligados entre si.

O tamanho de cada nó dimensionou-se de acordo com a métrica *out-degree*, o que permitiu identificar os utilizadores mais ativos na rede 3. Pela visualização do grafo, apresentado em (c), identificaram-se os três utilizadores mais ativos, PD, CG e TM, com valores de 49, 44 e 24 ligações, respetivamente, como apresentado em (b). Por definição, e como mencionado anteriormente, o *out-degree* representa o quão influente um nó pode ser dentro da rede. Estes resultados, para além de evidenciarem a actividade dos três utilizadores, demonstraram que eles eram os mais intervenientes dentro do grupo, ao responderem aos outros.

Na rede 3 calcularam-se o número de *likes* dos nós, conforme tabela (b) da Figura 37, visto que a sua importância não deve ser subestimada. A utilização de *likes*, quando aplicada aos *post*, como recurso linguístico, pode ser interpretada como uma posição a favor ou contra um ponto de vista submetido e dá uma ideia em tempo real (ou quase real) da aceitação do que está a ser publicado. Para além disso, os *likes* quando utilizados de forma

recíproca nos *posts*, podem ser interpretados como uma conversa assíncrona, que se desenrola durante um longo período temporal (dias ou até semanas). Este tipo de comunicação confirma que o conteúdo do discurso *web* reflete uma linguagem híbrida, composta por recursos linguísticos e não-linguísticos que se articulam entre si.

6.1.3 Processamento semântico

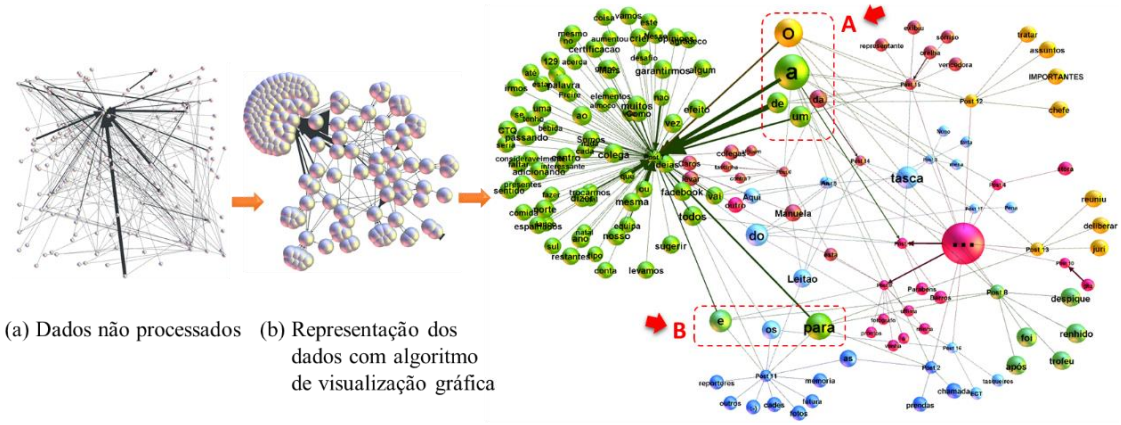
O processamento semântico implementado neste estudo de caso teve duas finalidades. A primeira, foi obter informação sobre os conceitos mais utilizados durante as trocas discursivas, para depois serem utilizados na análise das várias alternativas ligadas à organização do almoço do departamento. A segunda, foi identificar todos os aspetos ligados à complexidade do processamento semântico dos conteúdos publicados nas RSO. Para responder às duas questões, criou-se a rede semântica com todos os *posts* e respetivos conceitos neles contidos.

A extração e a estruturação dos conceitos contidos nos *posts* executaram-se com recurso ao *Excel*. Os dados obtidos nesse processo importaram-se para o *Gephi* para visualizar a rede em formato grafo. A representação gráfica da rede semântica permitiu identificar e ver padrões nos dados que de outra forma não seriam evidentes.

O processamento semântico do conteúdo dos *posts* teve como objetivo transformar dados qualitativos (não estruturados) em dados quantitativos. Este processamento tem aplicabilidade em praticamente qualquer tipo de comunicação escrita em linguagem natural, executando-se através da codificação das ocorrências dos conteúdos trocados. A transformação dos recursos linguísticos pode ser vista como o processo a partir do qual se fazem inferências e se tiram conclusões do que está contido nos *posts* das RSO.

Inicialmente visualizaram-se e exploraram-se os dados não processados como ilustra a Figura 38 (a). De seguida, interpretaram-se utilizando o *layout Force Atlas 3D* como mostra o grafo (b) da figura e, por último, avançou-se para o cálculo das métricas da SNA. O novo grafo configurado com os resultados obtidos nas métricas, assinalado como (c), foi utilizado como técnica de *text-mining* no apoio ao PLN.

Figura 38 - Visualização de dados para suporte ao *text-mining*

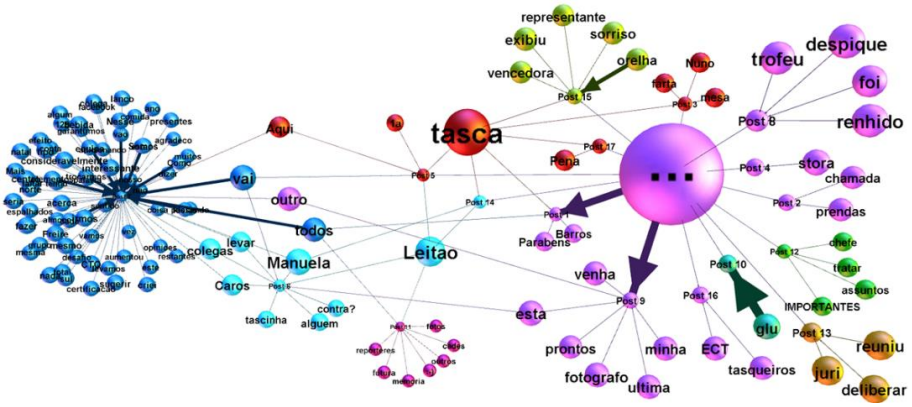


Como referido no capítulo 5, no PLN existem conceitos designados de *noise-words*, que fazem parte da linguagem natural e que podem ser descartados, pois não têm grande significado para a recuperação de informação. Através da visualização da Figura 38 (c) identificaram-se alguns conceitos irrelevantes tais como “para”, “de”, “a”, entre outros (exemplos assinalados em A e B). Após a limpeza das *noise-words* dos 17 *posts*, identificaram-se 122 conceitos que serviram de base para construção da matriz que liga e identifica que conceitos existem em cada *post*.

No início do processamento semântico, nos dados existiam um total de 259 conceitos. Após a eliminação das *noise-words* ficaram 122 conceitos, ou seja, descartaram-se 53% de dados irrelevantes. Com a sua eliminação, o texto ficou mais simples, a rede semântica menos densa, logo mais fácil de interpretar. Para efeito da análise, consideraram-se como conceitos quer as reticências (“…”) quer os *smiles* identificados nas trocas discursivas. Com base na matriz final construída, extraíram-se duas tabelas com os relacionamentos entre as entidades *post* e *concept*. As mesmas serviram de *input* para o *Gephi* e para a criação da rede 4.

Após o cálculo das métricas da SNA configurou-se o *layout* do grafo com os novos resultados, obtendo-se a Figura 39. A rede semântica (rede 4) continha todos os conceitos e *posts* identificados, ou seja, era composta por 155 ligações e 139 nós dos quais 122 eram conceitos e 17 eram *posts* com texto. A cor determinou-se utilizando a métrica *modularity class*, correspondendo cada cor a um grupo. O tamanho de cada nó corresponde à métrica *out-degree* para que os de maior dimensão identificassem os conceitos mais utilizados.

Figura 39 - Representação gráfica final da rede semântica



Através da análise visual da rede semântica apresentada na Figura 39, foi possível constatar que dos 17 *posts* analisados, só um tinha um número mais elevado de conceitos e os restantes tinham um número muito reduzido. Na figura, facilmente se podem identificar os diferentes *posts* e os conceitos que fazem parte de cada um. A aplicação do algoritmo da *modularity class*, no grafo, permitiu agregar todos os conceitos pertencentes a um mesmo *post*, de forma a que fossem interpretados como um subgrupo.

Pela visualização do grafo verificou-se também que as reticências (apesar de não serem uma palavra) foram as mais utilizadas. De uma forma geral, a linguagem utilizada no discurso *web* caracterizou-se pela tendência de reduzir a quantidade necessária de caracteres que se digitavam para expressar uma ideia e, consequentemente, para acelerar o processo de comunicação. Esse facto justifica o motivo pelo qual os utilizadores usaram esse recurso linguístico.

Através das métricas da SNA foi possível extrair uma lista das palavras-chave mais utilizadas e classificá-las. A Tabela 7 apresenta os cinco conceitos mais utilizados no discurso *web* analisado, bem como o número de *posts* em que foram empregues

Tabela 7 - Os 5 conceitos mais utilizados

Conceito	Quant. *Posts*	Nº de Vezes utilizado
...	14	18
tasca	6	6
Leitão	4	4
vai	3	4
todos	3	4

Analisar o discurso produzido nas RSO pode ser um risco, visto que não se sabe à partida que dados se vão conseguir extrair e qual o comportamento dos utilizadores. Para

além disso, é difícil predefinir a estrutura de uma qualquer RSO visto que, devido à sua dinâmica, está em constante mutação e evolução. De modo similar, não é possível predeterminar as suas fronteiras nem a natureza das suas ligações, visto que estas podem ser globais ou individuais, compactas ou dispersas.

Neste estudo de caso, em particular, o grupo analisado caracterizou-se por ser pouco ativo, por publicar poucos *posts* com texto (17 no total) e por colocar comentários com poucas palavras, o que justificou a baixa quantidade de palavras-chave encontradas no discurso. De referir que dos 50 *posts* publicados pelos utilizadores só 17 foram utilizados para construção da rede semântica. Os conteúdos dos restantes refletiam uma linguagem híbrida, mistura de fotografias e *likes*. Isto é, eram formados por recursos linguísticos e não-linguísticos que se articulavam entre si.

6.1.4 Conclusões do estudo de caso

O estudo de caso "Leitões" teve como objetivo perceber se o discurso produzido numa RSO (*Facebook*) podia ser utilizado para apoio à decisão e se as trocas discursivas podiam ser medidas através das métricas da SNA. No que se refere à utilização do discurso para apoio à decisão, concluiu-se que as diferentes métricas da SNA, das várias redes observadas, permitiram uma representação gráfica mais rica e mais estruturada dos intervenientes no discurso, assim como dos conceitos mais utilizados.

A interpretação dos dados, das quatro redes analisadas, com recurso às técnicas da SNA, visualização e métricas, permitiu uma melhor compreensão do problema e análise das alternativas disponíveis ligadas à organização do almoço do departamento. Permitiu também identificar, em cada rede, o utilizador mais importante do ponto de vista das métricas de centralidade da SNA.

Relativamente à análise das trocas discursivas, verificou-se que as mesmas não só podiam ser quantificadas e medidas como também o poderiam ser os conceitos existentes em cada uma delas. A extração de informação semântica com recurso às métricas da SNA, isto é, a quantificação do número de vezes que os conceitos surgem nos *posts*, permitiu determinar eventuais preferências e possíveis opiniões dos utilizadores. Além disso, a análise e visualização dos conceitos contidos no discurso *web* permitiram um acesso eficiente a informação chave. As palavras-chave identificadas, "tasca" e "leitão", permitiram perceber as preferências para o tipo de comida e espaço dos utilizadores envolvidos no discurso.

Os resultados obtidos, neste estudo de caso, mostraram que podiam ser utilizados para análise das várias alternativas do problema, bem como para apoio à tomada de decisão. Para além disso, permitiram concluir que o encadeamento das entidades do discurso *web*, todos os elementos do grupo e os respetivos interesses, ofereciam uma visão geral dos pontos de vista e da posição (a favor ou contra) do grupo.

Este estudo de caso foi o ponto de partida para a criação do modelo proposto, mostrando que era possível estruturar, de forma simples, as trocas discursivas, utilizando a SNA para extração de informação. Além do mais, concluiu-se que as técnicas utilizadas e os resultados obtidos na análise do discurso *web* podiam ser replicados noutras situações e utilizados nas fases do processo de decisão.

6.2 Estudo de caso - TAP

O estudo de caso "TAP", apresentado em Freire et al. (2015c), abordou a complexidade subjacente à análise de grandes volumes de dados, recolha e processamento de dados temporais (*snapshots*), focou os aspetos relevantes a considerar na sua análise, bem como a sua utilização no apoio à decisão. De referir, também, que foi através deste estudo de caso que se identificou a necessidade de incorporar no modelo um algoritmo para PLN. O algoritmo, para além de ter suportado as rotinas de limpeza dos dados não estruturados, transformou o conteúdo dos *posts* em tantos pedaços quantos os conceitos existentes.

Para efeitos de apoio à decisão, o foco da exploração e análise dos dados deste estudo de caso foi a satisfação do cliente e a identificação de conflitos durante a greve ocorrida em maio de 2015. Neste tipo de problema, as empresas normalmente utilizam pequenas amostras de dados e/ou recorrem a inquéritos para medirem e analisarem a insatisfação dos clientes em situações de crise. Os *call centers*, *e-mail*, *Skype* e balcões de atendimento, continuam a ser os canais de interação preferenciais para quem enfrenta estas situações. Todavia, as RSO são utilizadas cada vez mais pelos clientes, quando não há resposta através desses canais tradicionais.

Um dos desafios atuais é compreender o relacionamento dos clientes que comunicam *online*. A SNA, articulada com outras ferramentas, pode ser utilizada para alcançar esse entendimento e perceber como a comunicação entre eles e a empresa pode criar atrito e afetar a empresa. Embora não seja fácil capturar e processar o conteúdo textual, devido a especificidades da linguagem, a sua análise possibilita identificar clientes

insatisfeitos, identificando conceitos-chave por eles utilizados na comunicação *online*. Depois de identificar os clientes e caracterizar a insatisfação, as empresas podem decidir fazer algum tipo de intervenção. Por exemplo, efetuar um contacto telefónico mais personalizado, no sentido de explorar a fonte da insatisfação e resolvê-la. O problema de decisão em si, como identificar um cliente insatisfeito, é idêntico tanto nos métodos tradicionais de decisão, como nos atuais. O que muda é o acesso aos dados e as ferramentas disponíveis para auxiliar nas etapas do processo de decisão.

Esta greve de 10 dias foi caótica para a TAP e muitos clientes utilizaram o *Facebook* para reclamarem, confirmarem se tinha havido o cancelamento de um voo, cancelarem uma reserva, etc. O volume de dados recolhidos caracterizou-se por ser de maior dimensão, não só devido ao descontentamento dos clientes, mas também devido as características da companhia aérea. Apesar da greve ter decorrido entre o dia 1 e 10 de maio de 2015, recolheram-se também os dados do dia 11 e 12. Os dados da *FanPage* do *Facebook* da TAP recolheram-se de dois em dois dias para um melhor acompanhamento da situação e obtiveram-se seis conjuntos distintos de dados (*snapshots*).

6.2.1 Extração de dados

A extração de dados foi executada com o *NodeXL* e os dados apenas permitiram representar a entidade utilizador e as suas interações (*user* | *user*). Os *outputs* do *NodeXL*, em *Excel*, encontravam-se estruturados em formato grafo, isto é, com duas tabelas distintas. Uma que identificava todos os utilizadores da rede e outra que identificava todas as ligações entre eles.

A Tabela 8 caracteriza as recolhas de dados efetuadas entre o dia 1 e 12 de maio. A segunda coluna identifica o intervalo de recolha dos dados, as restantes colunas indicam o total de elementos das entidades *user*, *post* e *concept* identificados por *snapshot*. Quando comparámos os vários *snapshots*, observou-se que o primeiro, referente aos dias 1 e 2, registou o maior número de utilizadores (394). Observou-se, ainda, que o conjunto de dados dos dias 5 e 6 continham o maior número de *posts* e de conceitos (201 *posts* e 10.730 conceitos respetivamente). No total identificaram-se 1.328 utilizadores, recolheram-se 877 *posts* que continham 43.969 conceitos.

Tabela 8 - Tipo e quantidade de nós em cada conjunto de dados

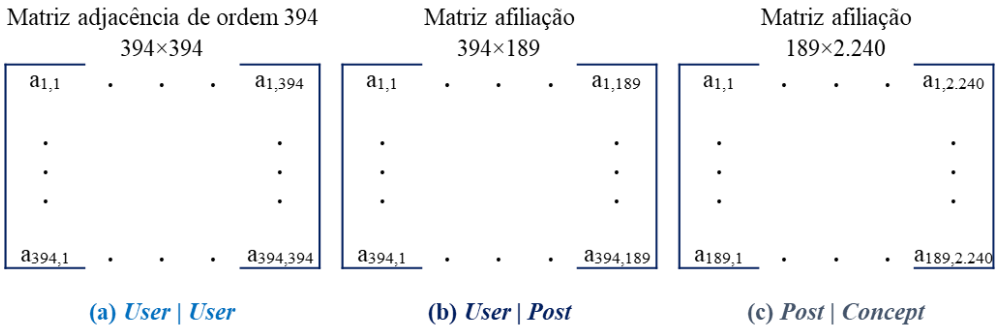
#	*Snapshot* (dias)	*User*	*Post*	*Concept*
1	01 e 02 de Maio	394	189	9.637
2	03 e 04 de Maio	216	143	7.909
3	05 e 06 de Maio	315	201	10.730
4	07 e 08 de Maio	168	144	6.741
5	09 e 10 de Maio	159	92	4.643
6	11 e 12 de Maio	76	108	4.309
	Total	**1.328**	**877**	**43.969**

6.2.2 Processamento e interpretação dos dados

O discurso produzido no *Facebook* possui pelo menos três entidades (*user*, *post* e *concept*) que podem ser analisadas individualmente ou de forma agregada. Como referido, o *NodeXL* apenas permitiu recolher dados das interações entre utilizadores. Assim, após o armazenamento no *Excel* dos dados originais, os mesmo foram processados e construídas mais duas redes. Para além da rede de utilizadores (*user* | *user*) já existente, para cada conjunto de dados, criou-se uma rede com as entidades utilizador e mensagem (*user* | *post*) e outra com as entidades mensagem e conceito (*post* | *concept*). As duas redes construíram-se utilizando o *Excel*. Uma permitiu identificar os *posts* mais relevantes e a outra identificar os conceitos mais utilizados.

Neste estudo de caso, para representar o discurso *web* de cada *snapshot*, foram utilizadas matrizes de adjacência e de afiliação. As matrizes de adjacência identificaram e representaram formalmente as ligações entre uma única entidade e as matrizes de afiliação caracterizaram as ligações entre cada uma delas. A Figura 40 exemplifica, utilizando informação do primeiro *snapshot*, as matrizes das três entidades do discurso *web*. A abordagem utilizada para agregar os três níveis da rede foi a transformação de redes *two-mode* numa única rede *one-mode*. O resultado final foi uma matriz única com as três entidades em análise (*user*, *post* e *concept*).

Figura 40 - Exemplo das matrizes do 1.º *snapshot*



(a) *User* | *User* **(b) *User* | *Post*** **(c) *Post* | *Concept***

O processamento inicial dos dados envolveu a criação de duas tabelas, no *Excel*, uma para armazenar os dados dos utilizadores e outra para armazenar o conteúdo das mensagens. Criaram-se *IDs* em ambas as tabelas, que permitiram não só identificar cada entidade do discurso *web* mas também caracterizá-la. Isto é, a estrutura da designação dos *ID*s dos utilizadores começava com o *user* seguindo-se uma numeração sequencial: "*user*-nnn". Para as mensagens foi seguida a mesma regra, isto é, "*post*-nnn". Foram ainda criados na tabela de *posts* atributos para identificar o tipo de nó e o idioma do *post*. Na tabela de utilizadores adicionou-se também um atributo para identificar o tipo de nó. Estes atributos permitiram relacionar os dados e a sua transformação do modelo relacional para o modelo *graph database*. O resultado foi um *output* de dados com os nós e as ligações entre as três entidades para analisar os três níveis de análise do discurso *web*.

Após o processamento e armazenamento na *graph database*, os dados foram explorados e interpretados no *Gephi*. O *Gephi* permitiu visualizar e manipular de forma simples as várias redes e aplicar técnicas de representação da SNA. As métricas da SNA permitiram identificar os nós mais relevantes de cada nível (entidade), classificá-los de acordo com a sua relevância e mostrar e descrever os resultados observados.

6.2.3 Processamento semântico

Em cada conjunto de dados selecionaram-se os *posts* que continham texto e criaram-se redes semânticas para identificação de palavras-chave. O conjunto de dados resultante, composto por 877 *posts*, foi processado com o *Excel*, para construir um *output* de dados em formato *graph database* para ser importado pelo *Gephi*.

Os *posts* recolhidos encontravam-se escritos em sete idiomas diferentes (DE: Alemão; ES: Espanhol; FR: Francês; IT: Italiano; NL: Holandês; PT: Português; UK: Inglês). O português e o inglês foram os idiomas mais representativos e mais utilizados, como ilustra a Tabela 9.

Tabela 9 - Percentagem de idiomas encontrados em cada conjunto de dados recolhidos

#	% Idiomas - *Post*								Total
	Snapshot (dias)	DE	ES	FR	IT	NL	PT	UK	*Posts*
1	01 e 02 de Maio		2%	4%			69%	26%	187
2	03 e 04 de Maio		2%	2%			59%	37%	136
3	05 e 06 de Maio	1%	5%	6%	0,5%		53%	36%	198
4	07 e 08 de Maio	0,7%	3%	6%	0,7%	1%	54%	35%	143
5	09 e 10 de Maio	2%	2%	7%	1%		60%	28%	92
6	11 e 12 de Maio	1%	5%	3%			66%	24%	74
	Total	**1%**	**3%**	**4%**	**0%**	**0%**	**60%**	**32%**	**830**

Os *posts* foram partidos em tantos pedaços quantos os conceitos existentes com recurso a um algoritmo criado para o efeito. O algoritmo criado para o processamento e extração de conceitos executava um conjunto de instruções de limpeza de dados. Os dados extraídos dos próprios *posts* alimentaram a *cleaning database* que auxiliou nas etapas de limpeza dos dados semânticos. A *cleaning database*, foi também alimentada com listas de *stopwords*, dos vários idiomas, encontradas na *web*[65]. Quer o algoritmo, quer a *cleaning database*, também foram utilizados para remover valores inválidos de alguns atributos e/ou uniformizar alguns dados. Por exemplo, datas, horas, número dos voos, designação dos aeroportos. No final de todo o processo de limpeza e transformação dos dados, construíram-se redes com os conceitos. O objetivo foi identificar palavras-chave e resumir o conteúdo dos *posts*.

Como se pode observar na Tabela 10, antes da limpeza e otimização do algoritmo e da *cleaning database*, os dados tinham na sua totalidade 43.969 conceitos. Após a eliminação e a limpeza dos dados irrelevantes, identificaram-se 22.624 conceitos válidos. Para esta análise, consideraram-se como conceitos as reticências ("..."), os *smiles* (":)"); ponto de exclamação ("!") e ponto de interrogação ("?") identificados nas trocas discursivas. No total descartaram-se 49% dos conceitos considerados irrelevantes e, apesar de eliminados, não retiraram sentido aos resumos finais obtidos.

Tabela 10 - *Noise-words* vs conceitos válidos

#	*Snapshot* (dias)	Valores absolutos			%	
		Noise-words	Conceitos válidos	Total conceitos	*Noise-words*	Conceitos válidos
1	01 e 02 de Maio	4.819	4.818	9.637	50%	50%
2	03 e 04 de Maio	3.673	4.236	7.909	46%	54%
3	05 e 06 de Maio	5.272	5.458	10.730	49%	51%
4	07 e 08 de Maio	3.241	3.500	6.741	48%	52%
5	09 e 10 de Maio	2.217	2.426	4.643	48%	52%
6	11 e 12 de Maio	2.123	2.186	4.309	49%	51%
	Total	**21.345**	**22.624**	**43.969**	**49%**	**51%**

6.2.4 Visualização gráfica

Após a construção das redes (uma para cada *snapshot*), utilizou-se o *Gephi* e a SNA para analisar as interações e a comunicação produzida pelos utilizadores. Os *layouts* das

[65] Exemplos: http://dev.mysql.com/doc/refman/5.7/en/fulltext-stopwords.html, https://www.docear.org/2012/09/28/list-of-6513-stop-words-for-17-languages-english-german-french-italian-and-many-others/. Último acesso em 7 de abril de 2020.

Figura 41, Figura 42 e Figura 43 ilustram as redes construídas na sequência da análise dos seis *snapshots* de acordo com três perspetivas diferentes. Na primeira perspetiva, analisou-se a estrutura das redes e as ligações entre as três entidades do discurso *web*. Na segunda, analisou-se a atividade e a comunicação entre os utilizadores e a TAP. Por último, a representação gráfica das redes semânticas permitiu analisar as mensagens, estruturar e interpretar os conteúdos semânticos dos *posts* e identificar palavras-chave.

Em cada *layout* indicou-se, no canto inferior esquerdo, a quantidade de nós e ligações entre eles. Combinaram-se de forma sistemática três algoritmos de agrupamento e *layout Force Atlas2* (Jacomy et al., 2014), *Fruchterman and Reingold* (Fruchterman e Reingold, 1991) e *Network Splitter 3D*[66] (Barão, 2014), para criar as representações gráficas das diferentes redes e para as comparar entre si. O primeiro algoritmo agrupa em torno de um nó principal, os nós diretamente ligados entre si, empurrando cada conjunto para longe uns dos outros (para destacar visualmente os diferentes subgrupos). O segundo, mantém mais próximos os nós diretamente interligados entre si. O terceiro, destaca as entidades da rede em camadas. Os dois primeiros algoritmos representam uma visão bidimensional da rede e o outro cria uma projeção tridimensional. A combinação destes três algoritmos ajudou na interpretação dos dados.

Comparativamente a outras fontes de informação tais como inquéritos, entrevistas, entre outros, analisar as RSO com recurso à SNA, para "ouvir" diretamente os clientes, permite decidir de forma mais célere em situações de crise. A análise da comunicação e das interações dos clientes da TAP, permitiram explorar e estruturar dados semânticos para extração de informação útil para responder a questões tais como: verificar o serviço e a capacidade de resposta da companhia aos clientes nos dias da greve; desenvolver indicadores utilizando a SNA, para rever e/ou reforçar as estratégias de atendimento ao cliente em situação de greve.

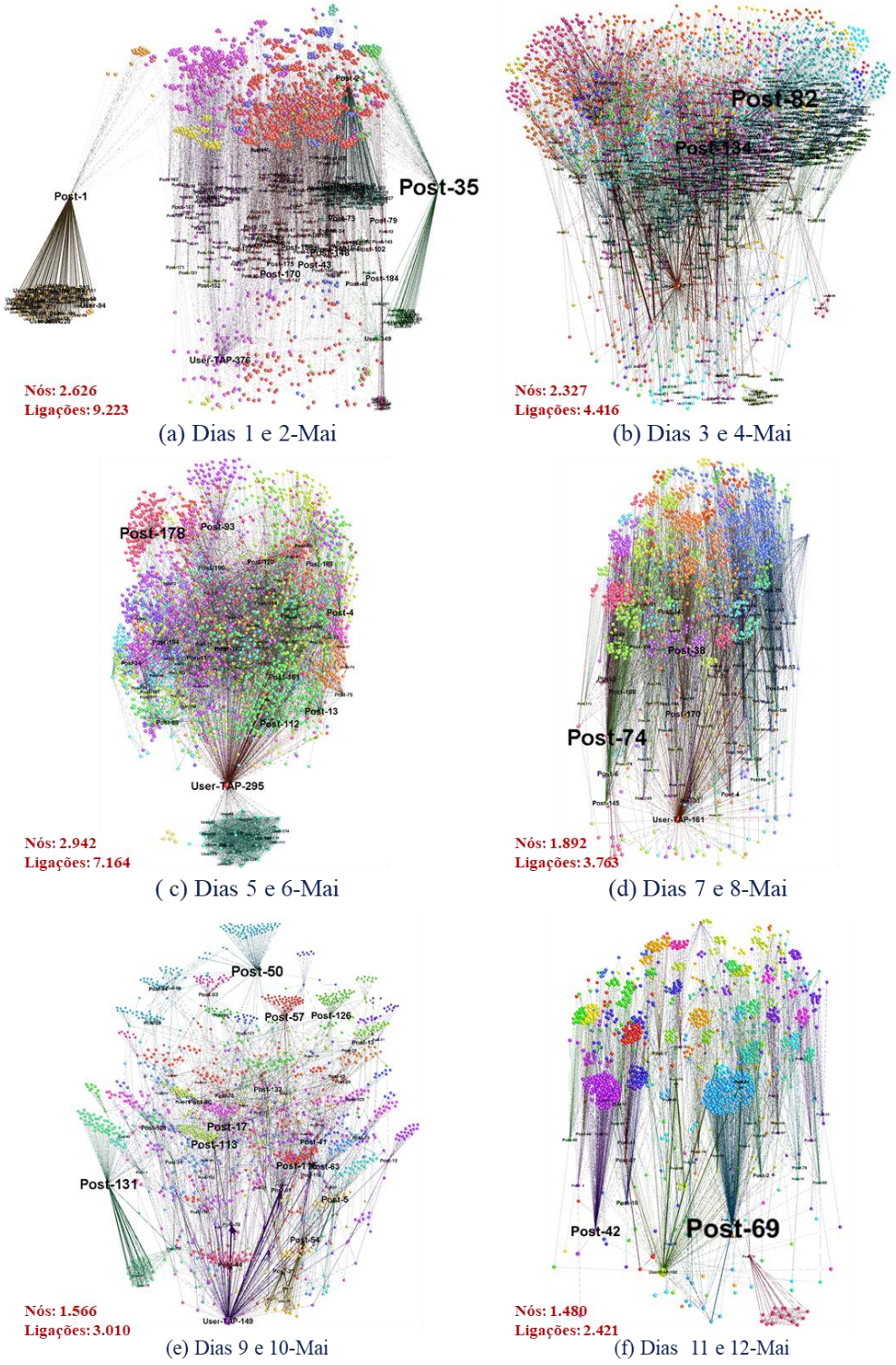
6.2.4.1 Rede global do discurso *web*

A Figura 41 ilustra as redes dos seis *snapshots,* após aplicação das métricas de centralidade mais importantes da SNA. As redes são compostas pelas três entidades do discurso *web*: *user, post* e *concept*. Como se pode observar, cada *layout* mostra os resultados obtidos em todos os intervalos de tempo analisados. Em cada *layout* o tamanho dos nós

[66] Algoritmo de visualização gráfica desenvolvido por Barão (2014) e incorporado nos *layouts* do *Gephi*.

alterou-se consoante as propriedades de cada rede e das características específicas das métrica utilizadas.

Figura 41 - Representação gráfica do discurso *web*



Para evidenciar e identificar os subgrupos da rede, utilizou-se a métrica *modularity class*. Como se observa nos grafos, as diferentes cores para além de caracterizarem os

subgrupos, representam os nós mais densamente ligados entre si. Utilizando o *Gephi*, calcularam-se também as métricas de centralidade da SNA tais como: *degree centrality*, *in-degree centrality*, *out-degree centrality*, *closeness centrality*, *betweenness centrality*, *PageRank* e *eigenvector*. Através das métricas de centralidade procurou-se identificar os nós mais importantes e como se posicionavam na rede. Para o efeito, aplicaram-se classificações numéricas para a aparência visual das redes, ao nível da dimensão dos nós e das suas legendas. O tamanho dos nós dimensionou-se com a métrica *betweenness centrality* e a dimensão da legenda caracterizou-se com a métrica *in-degree.* A primeira métrica indicou os utilizadores mais influentes na disseminação e no controlo da informação. Já a segunda, identificou os *posts* mais importantes, logo mais comentados ou visualizados.

A visualização dos *layouts*, permitiu ainda outras interpretações. Pelos *layouts* da Figura 41 observou-se que todas as redes tinham topologias e estruturas diferentes. Considere-se, por exemplo, a rede (a) da Figura 41 que representa o discurso *web* produzido nos dois primeiros dias de greve. Neste grafo, e de acordo com a *modularity class*, identificam-se dois subgrupos mais isolados comparativamente com outros existentes na rede (*post-1* na lateral esquerda e *post-35* na lateral direita da figura). Observa-se também que os grafos apresentados em (b), (c) e (d) respetivamente, eram as redes mais densas. Estas estruturas mais densas, compostas pelas ligações existentes entre nós, indicam a capacidade que os subgrupos têm em estabelecer ligações dentro e fora deles. Em sentido oposto, verifica-se que o grafo (e) era menos denso, o que permitiu de imediato identificar os subgrupos de menores dimensões.

As redes dos seis *snapshots* analisados, eram compostas por todos os utilizadores que tinham pelo menos uma interação com o utilizador TAP ou com qualquer outro utilizador da *FanPage* e com o conteúdo publicado na sequência das interações. As redes eram compostas pelas três entidades do discurso *web*: *user*, *post* e *concept*. A Tabela 11 resume os resultados quantitativos obtidos, em cada métrica, na análise dos seis *snapshots* para o utilizador TAP (dono da *FanPage*).

Tabela 11 - Métrica SNA do utilizador TAP

#	*Snapshot* (dias)	*In-degree*	*Out-degree*	*Closeness*	*Betweenness*	*PageRank*	*Eigenvector*
1	01 e 02 de Maio	92	277	0,777	0,008	0,008	0,262
2	03 e 04 de Maio	77	231	0,824	0,006	0,008	1,000
3	05 e 06 de Maio	130	369	0,887	0,010	0,009	0,680
4	07 e 08 de Maio	85	248	0,966	0,008	0,011	1,000
5	09 e 10 de Maio	51	162	0,833	0,006	0,007	1,000
6	11 e 12 de Maio	49	115	0,850	0,003	0,008	1,000

Por falta de capacidade de resposta dos canais de comunicação tradicionais, a *FanPage* da TAP foi utilizada pelos clientes durante a greve como *call center*. A Tabela 12 exemplifica algumas dessas mensagens publicadas pelos clientes.

Tabela 12 - Exemplos de mensagens

#	*Snapshot* (dias)	Mensagem	*Post_Id*	Idioma
1	01 e 02 de Maio	¿Qué pasa con los pasajeros de los vuelos que salen de Barcelona-Lisboa y están siendo cancelados ? TP1051 ¿Qué opciones se ofrecen ? ¿Se devuelve el dinero ?	Post-27	ES
		A tap tem os telefones desligados por causa da greve preciso de ligar para alterar viagens de 2 pessoas como fazer ?	Post-50	PT
2	03 e 04 de Maio	What happens when my flight is cancelled on Thursday? Do I have to book a flight on my own costs? Or do you take care of the costs?	Post-100	UK
		Por favor sabem informar alguma notícia do voo PRAIA-LISBOA TP1532 o meu voo é dia 4-05-2015 sabem informar algo se chego no aeroporto e não há voo como faço ... por favor estou com hotéis pagos em Portugal ... preciso uma posição	Post-11	PT
3	05 e 06 de Maio	Caros atendentes tenho um vôo marcado entre Salvador e Paris com conexão em Lisboa para o dia 13-05-2015 Tap0024. Já estará normalizado?	Post-6	PT
		i would like to confirm the flight TP1532 on 02-05-2015 which was cancelled due to the strike or not, if yes, what will be the rule for change the ticket?	Post-10	UK
4	07 e 08 de Maio	Ola Boa-tarde! Gostaria de saber se voo TP0331 se esta disponível porque não consigo fazer check-in obrigada	Post-63	PT
		Cara TAP, por favor consulte as minhas mensagens e responda ao meu pedido!!!! É a segunda vez que aqui publico hoje!	Post-79	PT
5	09 e 10 de Maio	Bom-dia, acabei de aterrar em Heathrow com a minha mulher vindos de Singapura e fui informado que o voo de ligação para Lisboa TP0363 das 11h20 está cancelado. Depois de 18h em aeroportos e aviões estamos exaustos e desejosos de chegar a casa. Devemos aguardar no aeroporto ou procurar alojamento? Obrigado.	Post-21	PT
		Bom-dia, o meu voo foi cancelado voo TP0808. Qual é a alternativa? O numero de confirmação do bilhete é 5ALVBG. Obrigado	Post-101	PT
6	11 e 12 de Maio	Gostaria de saber se é possível alterar a data de voos comprados em promoções com as tarifas mais baixas? Obrigada!	Post-101	PT
		Como é possível que continuem sem atender os telefones?!	Post-22	PT

Considere-se, por exemplo, o *snapshot* 3 dos dias 5 e 6 da Tabela 11, período em que se registaram os valores mais elevados nas métricas de centralidade *in-degree* (quantidade de *posts* ou comentários publicados na *FanPage*), *out-degree* (quantidade de respostas dadas pela TAP) e *betweenness centrality*. Equiparando o *Facebook* a um serviço de *call center*, pode assumir-se que a métrica *in-degree* do *user* TAP quantificou o número de chamadas recebidas e o *out-degree* quantificou o número de chamadas atendidas. Nessa perspetiva, seria possível calcular indicadores sustentados pela SNA sobre a capacidade de atendimento e resposta aos clientes.

Ainda na Tabela 11, o *snapshot* dos dias 7 e 8 evidenciou o valor mais elevado na métrica *closeness centrality* e *PageRank* (0,966 e 0,011 respetivamente). Uma das interpretações possíveis, para o valor da métrica *closeness centrality*, foi que esses foram os dias em que a TAP comunicou com mais utilizadores, através de um número mínimo de intermediários. Já o valor registado pelo *PageRank* evidenciou a importância da atividade da empresa com recurso à quantidade e qualidade das ligações que estabeleceu.

Relativamente à Tabela 13, para cada *snapshot*, mostram-se os utilizadores com valor mais elevado nas métricas de centralidade, comparativamente com todos os outros. De acordo com a tabela, verificou-se que a TAP (dona da *FanPage*) evidenciou os valores máximos nas métricas *in-degree*, *out-degree* e *betweenness centrality*. Estes resultados indicaram que a TAP tinha algum poder no controlo do fluxo de informação e era um utilizador ativo.

Tabela 13 - Utilizadores com valor máximo em cada métrica

#	*Snapshot* (dias)	*In-degree*	*Out-degree*	*Closeness*	*Betweenness*	*PageRank*	*Eigenvector*
1	01 e 02 de Maio	TAP	TAP	*	TAP	TAP	User-94
2	03 e 04 de Maio	TAP	TAP	*	TAP	User-58	TAP
3	05 e 06 de Maio	TAP	TAP	*	TAP	User-79	User-79
4	07 e 08 de Maio	TAP	TAP	*	TAP	TAP	TAP
5	09 e 10 de Maio	TAP	TAP	*	TAP	TAP	TAP
6	11 e 12 de Maio	TAP	TAP	*	TAP	TAP	TAP

No que se refere à métrica *closeness centrality*, verificou-se que, em cada intervalo de tempo, eram vários os utilizadores com o valor máximo de 1 (92, 83, 102, 58, 68 e 28 utilizadores respetivamente). Nenhum desses utilizadores era a TAP. O que indicou que, todos eles comunicavam mais do que a TAP com muitos outros, através de um número mínimo de intermediários. Logo, rapidamente faziam chegar informação a outros. Esta métrica está relacionada com o tempo que uma determinada informação demora a ser partilhada por todos os nós da rede. A sua interpretação em situações de crise é importante, por exemplo, para não permitir que um determinado cliente difunda rapidamente uma má notícia.

Relativamente à métrica *PageRank*, verificou-se que entre os dias 3 e 6 dois utilizadores, que não a TAP, registaram os valores mais elevados. Também na métrica *eigenvector* destacaram-se dois utilizadores que não a TAP, mas nos *snapshot* 1 e 3.

Para analisar as mensagens, entidade *post*, só foram utilizadas as métricas *in-degree*, *PageRank* e *eigenvector centrality*, conforme a Tabela 14. Os motivos que justificaram a utilização dessas três métricas foram vários. A métrica *in-degree*, por definição, quantifica as ligações recebidas. Nessa perspectiva, considerou-se que a métrica quantificava as ligações, num único sentido (*user* | *post*), estabelecidas de um utilizador para uma mensagem. Recorreu-se também às métricas *PageRank* e *eigenvector centrality*, pois se uma mede a importância de um nó, contabilizando a quantidade e a qualidade das ligações apontadas para ele, a outra considera que nós ligados a nós centrais são eles próprios centrais.

Tabela 14 - *Posts* com valor máximo em cada métrica

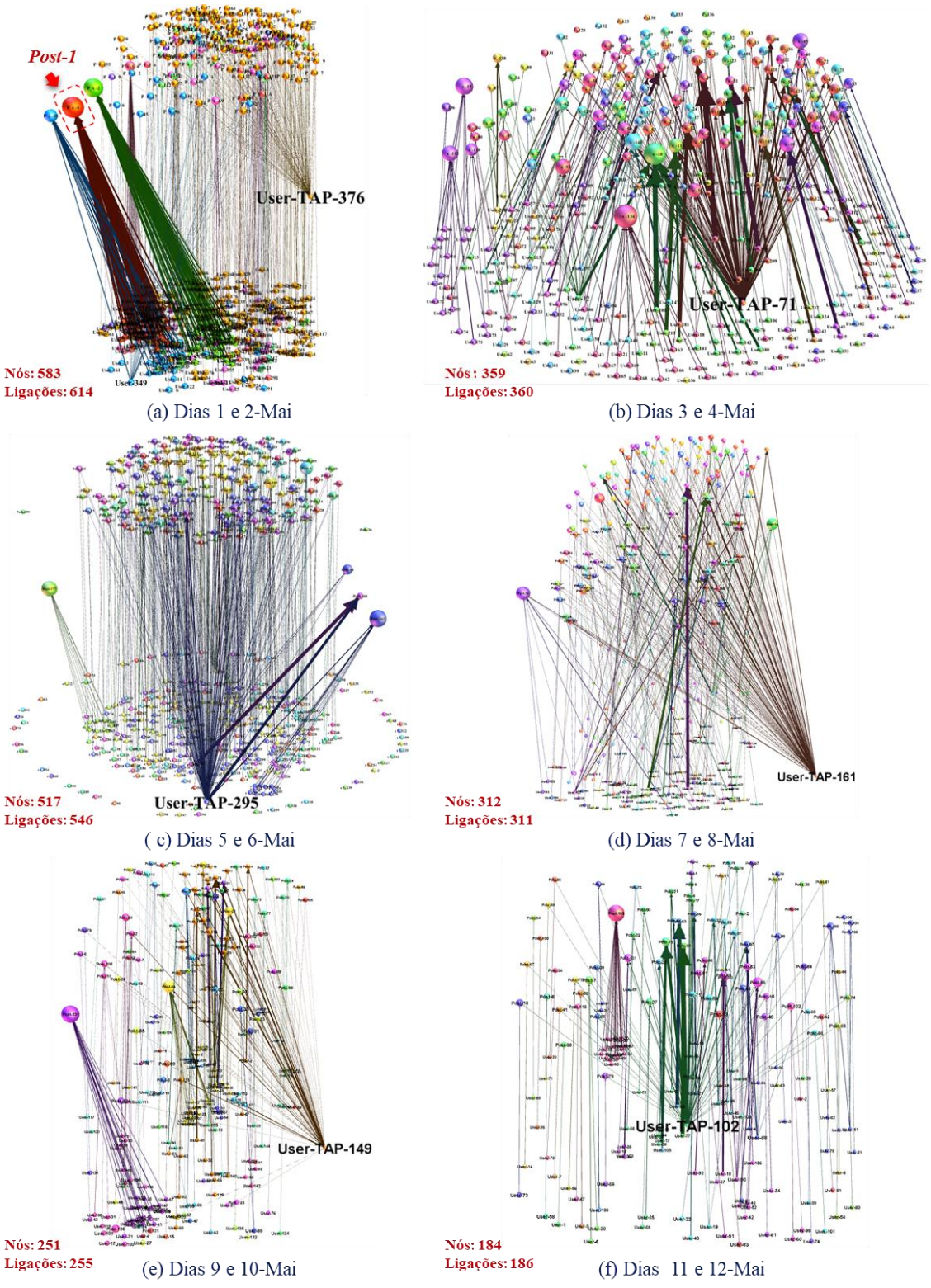
#	*Snapshot* (dias)	*In-degree*	*PageRank*	*Eigenvector*
1	01 e 02 de Maio	Post-35	Post-35	Post-1
2	03 e 04 de Maio	Post-82	Post-82	Post-82
3	05 e 06 de Maio	Post-178	Post-178	Post-4
4	07 e 08 de Maio	Post-74	Post-74	Post-74
5	09 e 10 de Maio	Post-50	Post-50	Post-21
6	11 e 12 de Maio	Post-69	Post-69	Post-69

Pela Tabela 14, observa-se que os resultados das métricas *in-degree* e *PageRank* identificaram os mesmos *posts* como sendo os mais importantes em todos os *snapshots*. Observa-se ainda que a métrica *eigenvector centrality* identificou *posts* diferentes em três dos *snapshots.* Independentemente disso, os resultados obtidos reforçaram a importância dos *posts* mencionados na tabela. Estas métricas conjugadas podem ser importantes para analisar qual ou quais as mensagens que estão a ser difundida rapidamente para outros clientes.

6.2.4.2 Rede *two-mode*: *user* | *post*

A Figura 42 ilustra os *snapshots* da janela temporal definida após aplicação das métricas de centralidade da SNA. Nesta análise desagregaram-se as redes e analisaram-se só duas das entidades do discurso *web*: os *users* e os *posts*. Como se pode observar, nos *layouts* o tamanho dos nós altera-se de acordo com as propriedades da rede e da métrica utilizada na representação. Para estas redes, os subgrupos identificaram-se com a métrica *modularity class*, a dimensão dos nós e das legendas definiu-se com a métrica *in-degree* e *out-degree* respetivamente.

Figura 42 - Rede *two-mode*: *user* | *post*



Utilizou-se a SNA para realçar os utilizadores mais relevantes, ao nível da sua posição estrutural na rede e ao nível dos conteúdos dos seus *posts*. Para verificar a capacidade

de resposta às questões colocadas pelos clientes na *Fanpage* da TAP, criaram-se redes compostas pelas entidades *user* e *post*. Para além disso, utilizaram-se as seguintes métricas:

- *In-degree* - Considerou-se que esta métrica era o número de contactos que o utilizador TAP recebeu. Por definição a métrica evidencia a quantidade de informação recebida bem como o número de ligações com outros nós, e mostra de forma clara a importância dos *posts* de cada utilizador. O facto de alguns *posts* serem muito comentados ou partilhados comprova a sua importância dentro da rede;
- *Out-degree* - Considerou-se que esta métrica era o número de respostas dadas pelo utilizador TAP, ou seja, a capacidade de resposta da companhia. Por definição o o*ut-degree* é uma métrica da influência de um nó dentro da rede e evidencia a capacidade que um utilizador tem em responder aos outros.

A visualização dos *layouts* da Figura 42 permitiu identificar nós mais isolados e sem ligação alguma com o utilizador TAP. Isto indicou que esses utilizadores (clientes) não tinham recebido resposta por parte da TAP, ao seus *posts*. Pela observação do *layout*, (Figura 42 (a)), e através de uma análise mais cuidada, sobressaíram três subgrupos associados a nós de grande dimensão. Considere-se, por exemplo, o subgrupo vermelho, ligado ao nó legendado como p*ost-1*. Para este caso verificou-se que a mensagem contida no *post*, nada tinha a ver com uma reclamação ou um pedido de informação. A mensagem colocada por um dos clientes (User-94) informava que, devido à greve, não poderia atuar nessa noite no clube de música eletrónica *Egg London*. Do ponto de vista da companhia, a identificação destas mensagens (*posts*) é importante (mesmo que a única decisão seja eliminá-las). A opinião *online* dos clientes tem vindo a transformar-se num recurso empresarial importante e este tipo de análise permite identificar informação útil, com o objetivo de limitar danos de imagem e evitar que um acontecimento evolua para uma crise.

Os nós que possuíam uma ligação estabelecida com o utilizador TAP eram os *posts* que foram "atendidos" por um operador e, por esse motivo, os clientes obtiveram uma resposta. Através da visualização dos grafos apresentados em (a), (b), (c), (d) e (e), foi possível verificar que a companhia, nos 10 dias de greve, respondeu a um número considerável de questões/reclamações colocadas pelos clientes no *Facebook*.

6.2.4.3 Rede semântica: *post | concept*

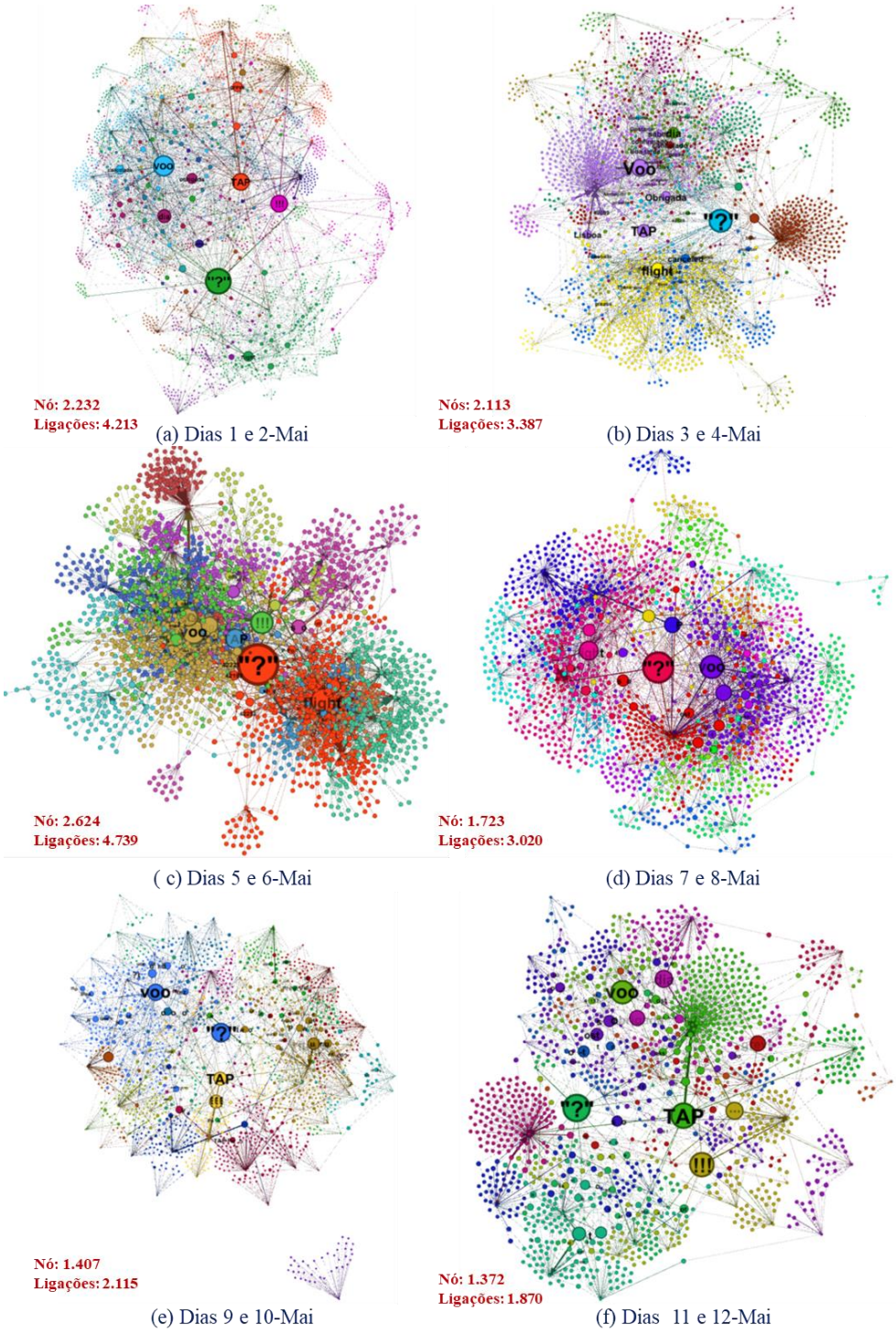
Considerando que as palavras-chave resumiam o texto contido nos *posts*, criaram-se redes semânticas com todos as mensagens e os conceitos a elas associadas. A extração de palavras-chave mostrou-se uma técnica importante para identificação dos conceitos mais relevantes utilizados no discurso *web*. Para o efeito selecionaram-se os *posts* contendo os recursos linguísticos das fontes de dados recolhidas e construíram-se as redes semânticas para cada uma delas (rede semântica: *post | concept*). O conjunto final de dados, composto por 769 *posts*, foi tratado com o *Excel* no sentido de obter dados para importar para o *Gephi*.

Os *layouts* da Figura 43 ilustram as redes semânticas construídas para analisar as mensagens e o seu conteúdo. As redes construíram-se com duas das entidades do discurso *web*: os *posts* e os *concepts*. A dimensão dos nós e das legendas definiu-se com a métrica *out-degree*. Com base na matriz inicial, construída para os seis *snapshots*, extraíram-se tabelas com os dados dos *posts* e com os conceitos, que serviram de *input* para o *Gephi* e para a criação das várias redes semânticas.

Partindo do pressuposto de que a análise do conteúdo semântico é um processo para extração de informação útil das mensagens trocadas entre utilizadores, considerou-se esse processo como a base a partir da qual se podiam tirar conclusões e fazer inferências sobre o que estava a ser dito. Uma análise textual permitiu a extração de uma lista das palavras-chave mais utilizadas e a sua classificação.

As redes de palavras-chave representadas nos *layouts*, para além de evidenciarem informação semântica relevante, também evidenciam a estrutura e a relação que existia entre os *posts* e os seus conceitos. Apesar de alguns dos *layouts* apresentarem semelhanças, quer em termos de densidade, quer em termos de estrutura, as métricas da SNA capturaram as diferenças escondidas e importantes numa análise semântica.

Figura 43 - Redes semânticas: *post* | *concept*



A Figura 44, para cada amostra de tempo, evidencia os vinte conceitos mais utilizados pelos clientes durante os 10 dias de greve. Para os identificar utilizou-se a métrica *out-degree*. Como exibido na figura, verificou-se que alguns sinais de pontuação, ponto de interrogação, ponto de exclamação e as reticências (apesar de não serem conceitos ou palavras) foram muito utilizados. O objetivo da utilização desses recursos linguísticos, por

parte dos clientes, foi para expressarem, de forma rápida, algumas emoções mais negativas de desagrado e de insatisfação associados à falta de respostas por parte da companhia.

Figura 44 - Os 20 conceitos mais utilizados em cada *snapshot*

01 e 02 de Mai-15

Conceito	*Out Degree*
"?"	97
voo	77
TAP	63
!!!	63
dia	48
obrigada	36
greve	34
flight	34
cancelado	30
Lisboa	26
...	22
voucher	21
muito	20
Empresa	19
thank-you	18
reserva	18
piloto	17
serviço	16
Saber	16
please	16

Posts = 189

03 e 04 de Mai-15

Conceito	*Out Degree*
"?"	80
voo	58
TAP	41
flight	40
dia	33
obrigada	27
!!!	27
Lisboa	24
canceled	23
saber	21
...	21
cancelado	19
confirmado	17
Boa-tarde	16
greve	14
gostaria	13
04/05/2015	13
please	12
RESPOSTA	11
Day	11

Posts = 143

05 e 06 de Mai-15

Conceito	*Out Degree*
"?"	112
voo	61
flight	61
!!!	56
TAP	48
dia	43
obrigada	40
Lisboa	31
canceled	31
greve	26
saber	24
...	21
thank-you	19
cancelado	19
Bom-dia	19
08/05/2015	18
07/05/2015	17
09/05/2015	16
please	15
TAP-Portugal	14

Posts = 201

07 e 08 de Mai-15

Conceito	*Out Degree*
"?"	74
voo	54
flight	43
dia	35
TAP	34
canceled	25
!!!	25
Lisboa	23
saber	21
obrigada	21
gostaria	19
thank-you	17
cancelado	16
09/05/2015	16
e-mail	15
please	14
viagem	14
Boa-tarde	14
Bom-dia	14
possível	13

Posts = 144

09 e 10 de Mai-15

Conceito	*Out Degree*
"?"	37
voo	32
TAP	25
!!!	25
flight	22
dia	18
obrigada	14
aeroporto	13
canceled	13
greve	12
please	11
cancelado	11
Lisboa	10
back	10
...	10
Bom-dia	10
saber	9
10/05/2015	9
thank-you	8
website	8

Posts = 92

11 e 12 de Mai-15

Conceito	*Out Degree*
"?"	25
TAP	23
!!!	22
voo	21
dia	15
...	15
obrigada	13
greve	13
flight	11
gostaria	10
saber	9
canceled	9
Boa-tarde	9
Lisboa	8
thank-you	7
Bom-dia	7
telefone	6
hotel	6
RESPOSTA	6
cliente	6

Posts = 108

Para comunicarem de forma mais rápida os utilizadores, das RSO, tendem a utilizar uma linguagem própria, caracterizada pela inserção e articulação de abreviaturas, *smiles*, imagens etc. Todavia, as trocas discursivas encontradas na *FanPage* da TAP, em nada se comparavam com as caracterizações e definições encontradas na literatura. As mensagens recolhidas caracterizaram-se por serem: muito extensas; formais e estruturadas (do tipo carta); tinham poucas abreviaturas ou outras formas de comunicação tais como *smiles*; apesar da insatisfação os clientes utilizaram uma linguagem cuidada. Curiosamente, verificou-se que os conceitos "Bom dia", Boa Tarde" e "Obrigada" foram muito utilizados independentemente da insatisfação. Para além disso, nos *posts* que registaram os valores mais elevados na métrica *in-degree*, isto é, as mensagens que foram as mais visualizadas ou comentadas em cada *snapshot*, verificou-se que ou tinham um elevado número de conceitos ou em sentido oposto um número reduzido. Por exemplo, o *post-134* dos dias 3 e 4, continha um total de 942 conceitos e em sentido oposto o *post-103* dos dias 11 e 12 só tinha 8

conceitos. A visualização destas duas mensagens[67] permitiu concluir que, apesar de ambas mostrarem descontentamento, se tratava de uma reclamação e que a outra denegria a imagem da companhia. Concluiu-se que a análise e a identificação dos conceitos podiam utilizar-se para o apoio à decisão. Relativamente ao *post-134*, a mensagem continha informação que poderia ser utilizada para responder ao cliente de forma assertiva, visto que os seus conceitos identificavam a hora, origem e destino do voo. No que se refere ao *post-103*, a sua identificação seria importante, mesmo que a única decisão fosse eliminá-la com o objetivo de limitar eventuais danos de imagem.

6.2.5 Conclusões do estudo de caso

O estudo de caso designado de "TAP" teve essencialmente dois objetivos. Em primeiro lugar, aplicar a SNA a conjuntos de dados de maior dimensão para identificar palavras-chave que resumissem o conteúdo do discurso *web* produzido nas RSO. Em segundo lugar, verificar se os conteúdos semânticos das trocas discursivas, recolhidos em intervalos de tempo regulares, poderiam ser utilizados para apoio à decisão em contexto de crise.

A utilização e o cálculo de métricas diferentes da SNA e a análise de múltiplas redes permitiram uma visão mais rica e estruturada dos envolvidos no discurso (TAP e seus clientes), bem como dos conceitos mais utilizados. A análise e a visualização dos conceitos podem ser utilizadas como *input* no acompanhamento e apoio de decisões, pois oferecem o acesso eficiente a informação importante. Os resultados obtidos na análise semântica oferecem ainda informação sobre as palavras-chave utilizadas durante as trocas discursivas, podendo ser utilizados para analisar várias alternativas de um problema e para apoiar a tomada de decisões.

A identificação de palavras-chave e a sua visualização gráfica, poderiam ser utilizadas para, em cada momento, reforçar a informação prestada aos clientes. Ou seja, um "olhar" regular e sistemático das trocas discursivas permitiria perceber quais as questões mais pertinentes e priorizar as respostas. Os resultados obtidos após a análise da *FanPage*, poderiam ser utilizados para responder aos clientes de forma mais assertiva, não só através desse canal de comunicação, mas de outros canais (*online*, balcões de serviço, etc.).

[67] Exemplo das mensagens disponível em: https://meocloud.pt/link/9e4e90c5-7065-4ec6-9286-b85b7b98c1c9/Extratexto/.

Permitiriam também, responder antes dos clientes fazerem perguntas. Por exemplo, a publicação antecipada de informação *online* tais como: se um voo vai ser ou não cancelado; a lista de voos cancelados ou disponíveis; como reagendar/alterar a data de um voo, entre outras, prestaria um serviço mais eficiente e evitaria descontentamento por parte do cliente.

Nos *posts* analisados, muitos clientes reclamavam que o *call center* não atendia as chamadas, ausência de respostas aos e-mails enviados, os balcões de atendimento dos aeroportos (assistência presencial) encontravam-se fechados e inexistência de informação sobre a greve no site da companhia. A visualização do discurso *web* auxiliaria na identificação de conversas relevantes e ofereceria informação direta e imediata sobre os sentimentos dos clientes. Permitiria ainda perceber a evolução da insatisfação do cliente e se estaria ou não a desenvolver-se negativamente e/ou a agravar-se. A análise e a visualização das trocas discursivas produzidas na *web* ajudariam na identificação de potenciais problemas de serviço, existentes nos canais tradicionais (*call center*, e-mail, etc.). A identificação desses problemas e a sua resolução atempada, permitiriam auxiliar em futuras ruturas e/ou falhas de serviço ocasionados por motivos de greve.

6.3 Estudo de caso - Morangos

O estudo de caso "Morangos", apresentado em Freire et al. (2017), teve como objetivo resolver um problema de tomada de decisão, utilizando o modelo proposto e descrito no capítulo 5, para extrair, processar, estruturar e analisar dados das RSO. De referir também que foi através deste estudo de caso que se identificou a necessidade de incluir no modelo a criação de redes semânticas para resumir o conteúdo dos *posts*. Como referido anteriormente, estas redes semânticas construíram-se estabelecendo uma relação entre os pares de conceitos (entidade *concept*) contidos num *post*.

No sentido de obter dados reais, optou-se por seguir uma ideia simples, vender morangos, e implementá-la junto de um grupo de potenciais clientes. Para o efeito, primeiro contactou-se um produtor de morangos com disponibilidade para entregar as encomendas no local de trabalho dos clientes. Depois, criou-se um grupo fechado na RSO *Facebook* e entregou-se uma caixa de morangos no local de trabalho dos clientes, à qual se anexou a seguinte mensagem:

> "Diretamente do produtor para o Cliente. Caso pretenda fazer encomendas adira ao grupo criado no *Facebook* Morangos e deixe o seu pedido".

À medida que os potenciais clientes aderiam ao grupo e faziam encomendas, em paralelo contabilizaram-se e registaram-se manualmente as quantidades encomendadas para posterior validação dos dados.

Por muito simples que possa parecer a ideia vender morangos, a mesma tinha um conjunto de questões associadas, tais como: validar a quantidade necessária de morangos para satisfazer as encomendas a entregar aos clientes; coordenar as entregas com os horários de saída do trabalho dos clientes; alterar a quantidade de morangos encomendadas ou novos pedidos que implicassem a verificação junto do fornecedor, para saber se ainda tinha morangos para entregar ou se já tinha realizado a expedição dos mesmos.

Assim, um dos principais objetivos foi prever a quantidade necessária de morangos para satisfazer as encomendas a entregar aos clientes. Nesse sentido, descrevem-se algumas questões associadas ao problema de decisão definido.

Primeiro, era difícil prever com antecedência os pedidos submetidos e saber quando surgiria uma nova mensagem (*post*) no grupo para responder rapidamente. Não era realista, responder e aceitar um pedido no dia seguinte, com base em informação do dia anterior (a não ser que o pedido fosse nesse sentido). Portanto, as encomendas aceitaram-se e entregaram-se sempre no dia em que o pedido era submetido (através de mensagem no *Facebook*).

Segundo, era difícil prever exatamente a quantidade de morangos que cada cliente iria encomendar diariamente. Logo, considerou-se existir uma margem de erro nas quantidades encomendadas, isto é, uma variação para cima, para baixo, ou nula. Para este exemplo de aplicação, essa margem permitiu validar se a quantidade encomendada ficava acima ou abaixo do realmente encomendado.

Terceiro, como era difícil prever grandes variações num grupo grande e disperso de clientes, utilizou-se um conjunto de pessoas sediadas no mesmo local durante o seu período laboral. Isso permitiu restringir o volume de trocas discursivas dentro do grupo e reduzir o tamanho dos *posts*. Assim, e de forma análoga ao estudo de caso "Leitões", garantiu-se que eram criados menos eventos no *Facebook* e a utilização de uma rede mais pequena permitiu limitar o volume de dados e a quantidade de elementos em análise.

Por último, era difícil associar um "pedaço de texto" de um *post* a uma alteração da quantidade encomendada. À partida, qualquer *post* podia influenciar a quantidade total encomendada. A aceitação dessa premissa criava um problema para resolver. Como

selecionar os *posts* mais relevantes, entre todos os inseridos diariamente? Para ultrapassar essa questão, na análise semântica, consideraram-se apenas os *posts* recebidos dos clientes e não os enviados pelo administrador do grupo. Isto implicou que existia um grau de imprecisão, pois o risco de não considerar um *post* com uma encomenda afetava a quantidade final a ser entregue. Mas, considerou-se uma simplificação aceitável.

6.3.1 Extração de dados

Os dados utilizados na análise recolheram-se ao longo de uma semana com recurso ao *Netvizz* que, de forma simples e rápida, permitiu importar os dados do grupo. Como exemplificado na Tabela 15, extraíram-se cinco fontes de dados com a atividade dos utilizadores em torno dos *posts* do grupo.

Tabela 15 - Descrição das fontes de dados recolhidas

#	*Output* de dados	Formato	Características
1	group_AAAA_MM_DD_hh_mm_ss.gdf	GDF	Ficheiro com rede *two-mode* em formato grafo. Os nós eram *posts* ou utilizadores.
2	group_AAAA_MM_DD_hh_mm_ss_interactions.gdf	GDF	Ficheiro com rede de interações entre utilizadores, em formato grafo. Os nós eram utilizadores.
3	group_AAAA_MM_DD_hh_mm_ss.tab	TSV	Ficheiro em formato TSV (texto). As linhas correspondiam aos *posts*.
4	group_AAAA_MM_DD_hh_mm_ss_statsperday.tab	TSV	Ficheiro com resumo estatístico por dia de: *posts*; *likes*; *comments* e partilhas.
5	group_AAAA_MM_DD_hh_mm_ss_comments.tab	TSV	Ficheiro em formato TSV (texto). As linhas correspondiam aos *comments*.

AAAA_MM_DD_hh_mm_ss: data e hora do ficheiro

Os cinco *outputs* recolhidos continham dados de dois tipos: estruturados e semiestruturados. Os dados estruturados, dos ficheiros 1 e 2, encontravam-se em formato grafo. O ficheiro 1 continha dados das entidades *user* e *post* que permitiram criar uma rede *two-mode*. Os dados deste ficheiro caracterizavam as interações entre os utilizadores e os respetivos *posts*. Já o ficheiro 2 continha a rede de interações entre utilizadores, ou seja, só continha os nós da entidade *user*.

Os dados semiestruturados, ficheiro 3 com os *posts* e ficheiro 5 com respetivas respostas, encontravam-se em formato texto. No ficheiro 3 as linhas correspondiam aos *posts* trocados entre utilizadores. Já no ficheiro 5 as linhas correspondiam às respostas colocadas nos *posts*. Na fase de processamento, o encadeamento das ligações destes dois conjuntos de dados executou-se com recurso a técnicas de *data-mining* relacionando os *IDPost* e *IDUser*.

Com os *outputs* de dados dos ficheiros 1 e 2, criaram-se duas redes distintas:

- Rede 1: Contendo todos os utilizadores com interação entre eles e respetivos *posts*. Esta rede, *user* | *post*, era constituída por 46 nós e 71 ligações entre eles;
- Rede 2: Contendo todos os utilizadores com interação. Esta rede, *user* | *user*, era composta por 13 nós e 48 ligações entre eles.

Para a análise semântica utilizaram-se os conteúdos dos *posts*, capturados como texto simples dos ficheiros 3 e 5. Desta forma garantiu-se a distinção entre o conteúdo textual discursivo, relevante para a análise, e as outras formas de linguagem utilizadas para comunicar. Estes dados extraídos dos *posts* caracterizavam-se por serem dados não estruturados, o que implicou o seu processamento e estruturação.

6.3.2 Processamento e interpretação de dados

Após a extração dos dados, interpretaram-se as Rede 1 e Rede 2, para identificar eventuais dados irrelevantes. Numa primeira fase, porque estamos a falar de dados sociais, o objetivo foi criar um conjunto de técnicas para explorar os dados para obter informação útil para o apoio à decisão organizacional. Para o efeito recorreu-se ao *data-mining* através da SNA. Os dados originais, sem qualquer processamento, inseriram-se no *Gephi*, para identificar *posts* irrelevantes.

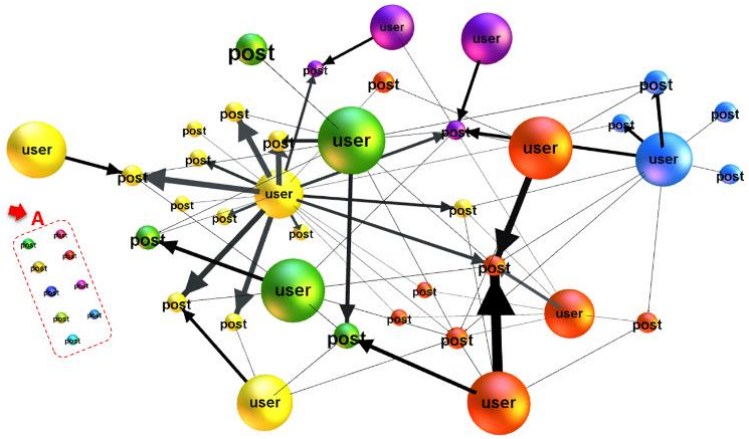
A utilização do *Gephi* permitiu produzir automaticamente representações gráficas das redes para identificar subconjuntos específicos de nós, através da atribuição de cores e legendas. A visualização dos grafos simplificou o processo de captar e interpretar a informação com base nas características dos dados. Para além disso, o *Gephi* permitiu calcular as métricas da SNA para caracterizar o papel que os utilizadores desempenhavam dentro do grupo. Os resultados foram sendo utilizados em testes exploratórios, para assim analisar e ajustar os dados de acordo com as necessidades.

Utilizou-se a métrica *modularity class* para identificar os subgrupos, onde cada cor representava um grupo. Para os diferentes tamanhos dos nós, que refletiam as diferenças na participação, utilizou-se a métrica *closeness centrality*. Esta métrica representa independência e evidenciou quem comunicava com outros utilizadores do grupo, através de um número mínimo de intermediários. Nos grafos das várias redes, as diferentes espessuras das linhas representavam o fluxo de interações entre as entidades.

A visualização e interpretação da Figura 45 e da Figura 46, permitiu tirar algumas conclusões. Como ilustra a Figura 45 na Rede 1, construída com os dados originais,

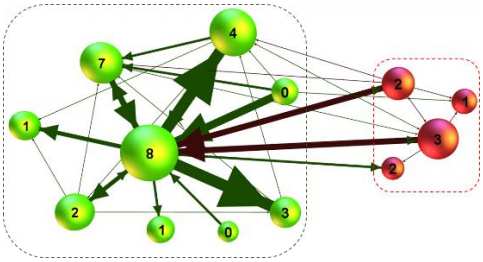
identificaram-se, no lado esquerdo, *posts* sem ligação alguma a outro *post* ou a um utilizador. Os mesmos foram considerados irrelevantes para a análise e eliminados, pois eram mensagens automáticas do sistema colocadas pelo próprio *Facebook*. Como se pode observar pelo grafo apresentado na figura, identificaram-se vários grupos distintos (amarelo, verde, laranja, azul, etc.). Visualmente conseguiu associar-se a entidade *post* à entidade *user*. Para identificar as encomendas efetuadas por cada utilizador, na fase do processamento dos dados, atribuíram-se os nomes dos clientes aos respetivos nós. Assim, identificaram-se as pessoas e os *posts* que faziam parte de cada subgrupo e qual era o mais influente.

Figura 45 - Rede 1: *users* e *posts* (dados originais)



Na Rede 2 da Figura 46, também construída com os dados originais da interação entre utilizadores, identificaram-se dois subgrupos distintos, assinalados a verde e vermelho. A espessura das linhas representa o número de vezes que um utilizador comentou ou gostou do *post* de outro.

Figura 46 - Rede 2: interação entre *users* (dados originais)



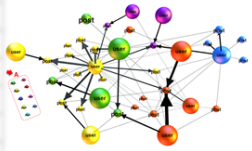
Nesta etapa de processamento alteraram-se os modelos de dados originais, como indicado na Figura 47 e na Figura 48. Para além da alteração dos atributos nas tabelas de nós e ligações das redes, no *Excel* criaram-se mais duas tabelas. Uma para armazenar a informação da entidade *user* e outra para armazenar os *posts* e os seus comentários.

Atribuíram-se *IDs* únicos a todos os registos, formados pela designação *user*, *post* ou *comment* seguidos de uma numeração sequencial.

Figura 47 - Modelo de dados pré e pós processamento (rede 1 e rede 3)

Rede 1: atributos

Nós	Ligações
Id	Source
Label	Target
type	Type
type_post	Id
post_by_id	Label
post_by_name	Weight
post_published	Community
post_published_unix	Neighborhood Overlap
likes	Embeddedness
likes_count_fb	
comments_count_fb	
comments_all	
comments_base	
comments_replies	
comment_likes	
shares	
engagement	
post_id	
post_link	



(a) Ficheiro formato grafo (*two-mode*)
Dados originais

Rede 3: atributos

Nós	Ligações
Post_id num	Source
Node	Target
ID_Novo	Type
Tipo_Node	Node
Data_post_published	Tipo_Rede
Is Eliminar?	Is Eliminar?
likes	
likes_count_fb	
comments_count_fb	
comments_all	
comments_base	
comments_replies	
comment_likes	
shares	
engagement	

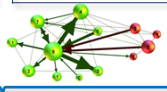
(b) Ficheiro formato grafo (*two-mode*)
Dados processados

Como se pode observar nas Figura 47 e Figura 48, em que (a) representa as tabelas e os atributos dos dados originais e (b) representa os atributos após o processamento e armazenamento na *graph database*. Nesta base de dados representou-se cada rede com duas tabelas, uma com os registos dos nós e outra com os registos das ligações entre os vários nós. Tecnicamente a *graph database* foi o processo definido para armazenar os dados em formato grafo dos nós, ligações e interações sociais.

Figura 48 - Modelo de dados pré e pós processamento (rede 2 e rede 4)

Rede 2: atributos

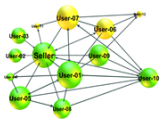
Nós	Ligações
Id	Source
Label	Target
actions	Type
made_posts	Id
made_comments	Label
made_likes	Weight
rec_comments	Community
rec_likes	Neighborhood Overlap
rec_commentlikes	Embeddedness



(a) Ficheiro formato grafo (interações)
Dados originais

Rede 4: atributos

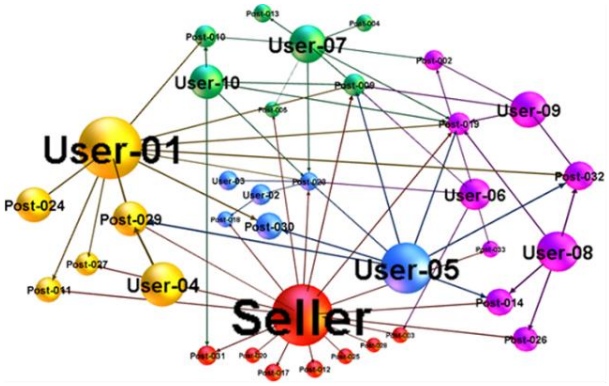
Nós	Ligações
User_Acronimo	Source
ID_Novo	Target
actions	Type
made_posts	Node
made_comments	Tipo_Rede
made_likes	
rec_comments	
rec_likes	
rec_commentlikes	



(b) Ficheiro formato grafo (interações)
Dados processados

Após a eliminação dos *posts* irrelevantes, os dados inseriram-se novamente no *Gephi* e construiu-se a rede 3, Figura 49 (a). Esta nova rede foi construída com todos os elementos do grupo, que tinham pelo menos uma ligação a um *post*, tendo a rede ficado com 36 nós e 67 ligações. A tabela apresentada em (b), na Figura 49, ilustra os resultados obtidos na análise da rede. Observou-se que o utilizador mais ativo foi o *User-01*, porque registou um *out-degree* superior aos outros (10). Isso significa que criou vínculos entre 10 outros utilizadores através de *likes* ou adicionou novos *posts*. Em sentido oposto, o utilizador *User-03* foi o elemento com menor valor na métrica *out-degree* (1). Isso significou que só escreveu um *post*. Os utilizadores com valores mais elevados nesta métrica evidenciam mais ligações e tendem a ser mais poderosos, porque afetam de forma direta os outros utilizadores. Através do grafo (a) da Figura 49, de forma rápida e simples, a dimensão dos nós evidenciou a importância que os utilizadores acima referidos tinham dentro do grupo.

Figura 49 - Visualização e resultados da rede 3 após *data-mining*



(a) Rede 3: *user* e *post*

Utilizador	Subgrupo	*Out-degree*
Seller	0	19
User-01	1	10
User-04		1
User-07	2	8
User-10		5
User-05	3	7
User-02		2
User-03		1
User-06	4	6
User-08		4
User-09		4

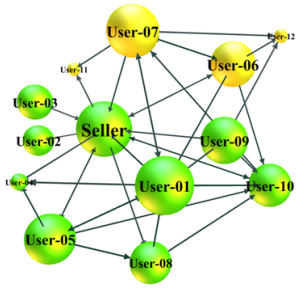
(b) Métricas SNA

A rede 4, (a) da Figura 50, obteve-se após o processamento dos dados da rede 2 e era constituída por todos os elementos do grupo que tinham pelo menos uma ligação a outro qualquer elemento. Enquanto que as duas redes anteriores caracterizavam-se por serem redes *two-mode*, com utilizadores e *posts*, a rede 4 caracterizava-se por ser *one-mode* e possuir 13 nós e 48 ligações. A representação gráfica, otimizada com as métricas de centralidade, permitiu avaliar as ligações entre utilizadores, identificar as propriedades de interação mais significativas, nomeadamente em termos de importância, poder, prestígio, influência e difusão de informações dentro da rede.

Pela análise dos resultados obtidos, resumidos na tabela (b) da Figura 50, observou-se que os elementos mais ativos foram os *User-01*, *User-07* e *User-05*, porque registaram um valor na métrica *out-degree* superior aos outros (6). Isso significou que

criaram vínculos entre 6 outros utilizadores. O *out-degree* é uma métrica da influência de um utilizador dentro da rede, evidenciando a sua capacidade de resposta aos outros. Neste contexto de negócio, isso foi importante porque os clientes (vistos como utilizadores) com valores elevados de *out-degree*, logo com maior poder na RSO, influenciavam os outros.

Figura 50 - Visualização e resultados da rede 4 após *data-mining*



(a) Rede 4: interação entre utilizadores

Utilizador	*Degree*	*In-Degree*	*Out-Degree*	*Closeness*	*Betweenness*	*PageRank*	*Eigenvector*
Seller	19	9	10	0,91	0,33	0,12	0,98
User-01	**14**	**8**	**6**	**0,67**	**0,08**	0,10	**1,00**
User-07	11	5	**6**	**0,67**	0,07	0,08	0,81
User-10	11	7	4	0,59	0,03	0,09	0,84
User-05	9	3	**6**	0,63	0,00	0,05	0,51
User-06	7	2	5	**0,67**	0,02	0,04	0,37
User-08	6	3	3	0,45	0,00	0,05	0,38
User-04	5	4	1	0,00	0,00	**0,32**	0,64
User-02	4	2	2	0,50	0,00	0,05	0,26
User-09	4	0	4	0,61	0,00	0,02	0,00
User-12	3	3	0	0,00	0,00	0,04	0,24
User-11	2	2	0	0,00	0,00	0,04	0,37
User-03	1	0	1	0,50	0,00	0,02	0,00

(b) Métricas de centralidade da SNA

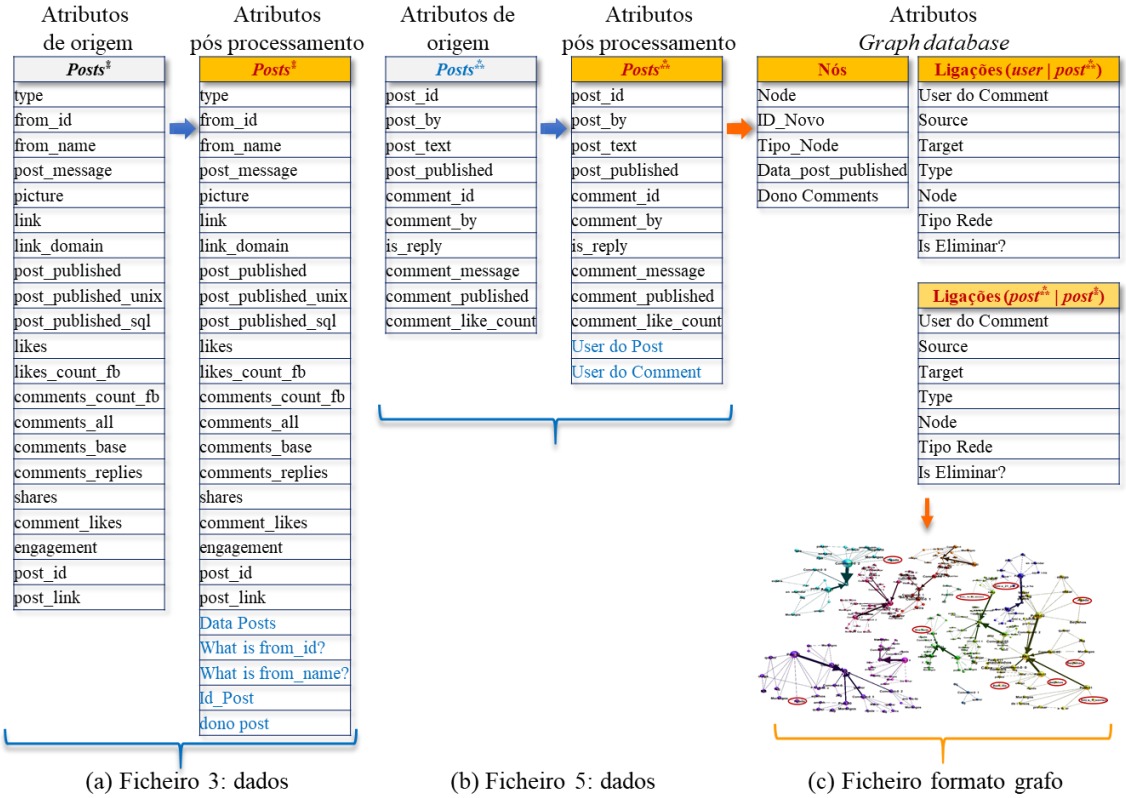
De todos os utilizadores, o *User-01* foi o elemento com valor mais elevado na métrica *in-degree* (8), o que significa que os seus *posts* foram os mais visualizados ou comentados. O facto desses *posts* terem sido considerados mais interessantes e/ou relevantes pelo grupo, mostra claramente a importância desse utilizador.

De um modo geral, o *User-01* evidenciou valores elevados em todas as métricas. Registou também os valores mais elevados nas métricas *closeness centrality*, *betweenness centrality* e *eigenvector centrality* (0,67; 0,08 e 1 respetivamente). O facto de ter um valor elevado de *closeness centrality* representa independência e que comunicava com os outros elementos através de um número mínimo de intermediários. Os resultados elevados do *eigenvector centrality* confirmam que o *User-01* era um dos utilizadores mais importantes, visto que esta métrica tem como objetivo medir a importância de um nó, em função da importância dos seus vizinhos.

Estas métricas da SNA revelaram algum nível de poder do utilizador *User-01* designado de utilizador-chave e forneceram informação relevante para apoio à decisão, pois ele influenciou a quantidade de morangos que os outros encomendaram. Através da Figura 50 (a), visualiza-se de forma rápida e simples a importância que os utilizadores, acima referidos, tiveram dentro do grupo.

Ainda na fase de processamento, estruturaram-se os dados dos ficheiros 3 e 5 que continham os *posts* e os seus comentários. A Figura 51 representa o modelo de dados antes e depois do processamento dos *posts* e dos *comments*. Estes dados armazenaram-se em tabelas diferentes, uma de *posts* e outra de *comments*. A sua estruturação implicou estabelecer uma relação entre ambas as tabelas, recorrendo aos atributos criados para identificar quem comentou ou publicou o *post*. Por último, os dados estruturados foram transformados e armazenados na *graph database*.

Figura 51 - Modelo de dados pré e pós processamento (*posts*)



(a) Ficheiro 3: dados *posts* iniciais*

(b) Ficheiro 5: dados *posts* resposta**

(c) Ficheiro formato grafo *user* | *post*** e *post*** | *post**

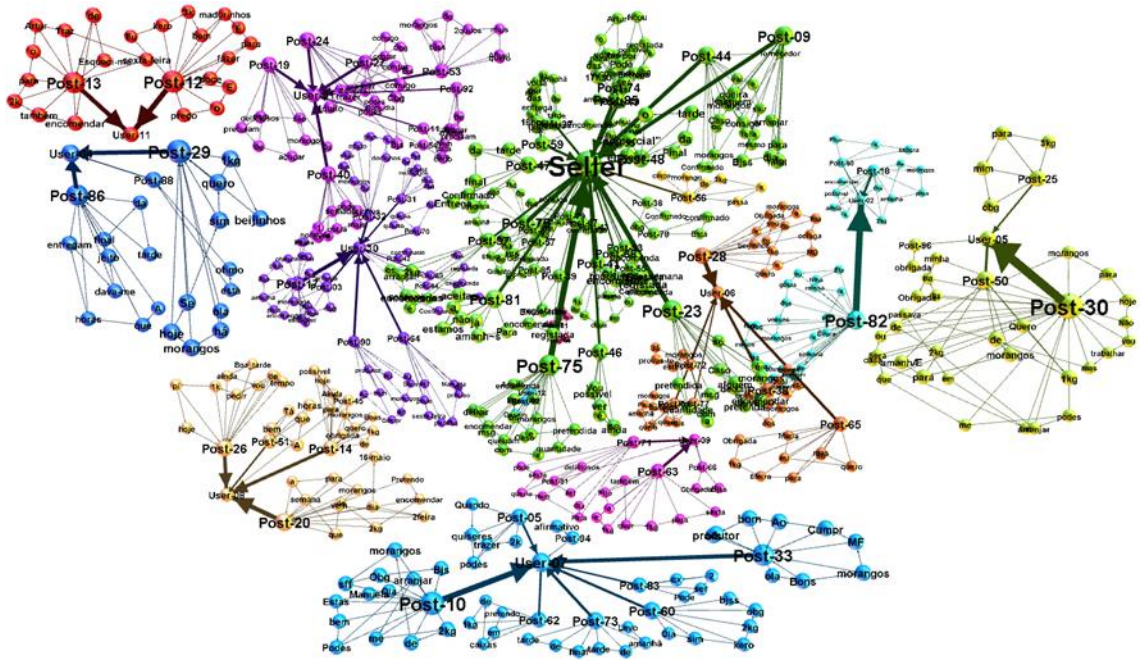
Após analisar as redes anteriores, e com o objetivo de obter um resumo das trocas discursivas, criou-se a Rede 5. A abordagem utilizada para agregar os três níveis da rede foi a transformação de redes *two-mode* numa única rede *one-mode*. O resultado final foi uma rede única com as três entidades em análise (*user*, *post* e *concept*). A criação desta rede envolveu a agregação dos dados dos utilizadores, *posts* e conteúdo textual dos *posts*, como definido no modelo apresentado no capítulo 5.

Para extração dos conceitos dos *posts* utilizou-se a *cleaning database* e o algoritmo criado para o efeito. A *cleaning database* alimentou o algoritmo de limpeza e de

padronização, utilizando tabelas de *noise-words*, *smiles*, sinónimos e pontuação, que auxiliaram o processamento dos dados não estruturados. A *cleaning database* utilizou-se para: extrair a rede semântica do discurso *web*; extrair uma rede semântica para perceber a relação entre conceitos contidos no texto das trocas discursivas; limpar e uniformizar o texto, por exemplo através da remoção de espaços e eliminação de palavras desnecessárias.

Inicialmente, os dados foram inseridos no *Gephi* sem qualquer processamento, para identificar conceitos irrelevantes. A sua identificação foi um processo recursivo e sistemático executado com recurso a técnicas de visualização. A rede 5, representada na Figura 52, obtida com os dados originais era muito mais densa e apresentava uma topologia diferente do momento inicial para o final.

Figura 52 - Rede 5: exemplo de visualização dos dados para apoio ao *text-mining*



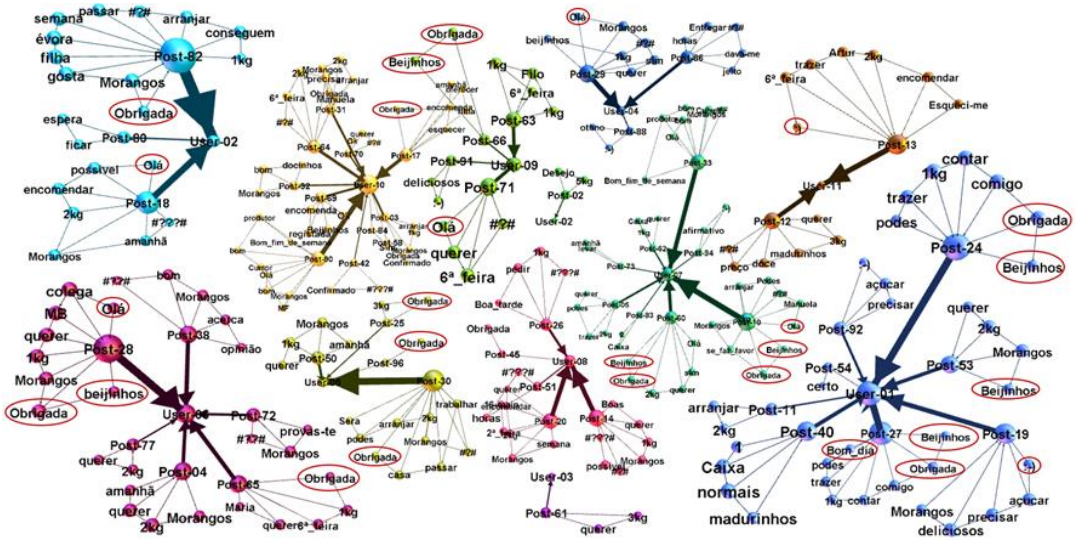
Como se pode observar pela Tabela 16, antes da otimização da *cleaning database*, a rede continha 673 nós e 1.148 ligações. Numa segunda etapa, após limpeza de alguns dados irrelevantes, a rede passou a ter 309 nós e 480 ligações. Inicialmente as mensagens continham 574 conceitos e, após a eliminação dos dados irrelevantes, restaram 171. Descartaram-se 70% dos conceitos por terem sido considerados irrelevantes e, apesar de eliminados, não retiraram sentido aos resumos finais obtidos.

Tabela 16 - Tipo e número de nós e ligações em cada conjunto de dados

	Entidades	Dados originais	Dados intermédios	Dados finais	% Dados finais
(a) Tipo e número de nós em cada rede	*User*	13	13	13	100%
	Post	86	56	53	62%
	Concept	574	240	171	30%
	Total nós	**673**	**309**	**237**	**35%**
(b) Tipo e número de ligações em cada rede	*Post | User*	86	56	53	62%
	Concept | Post	574	240	171	30%
	Concept | Concept	488	184	118	24%
	Total ligações por entidade	**1.148**	**480**	**342**	**30%**

A criação e visualização da rede 6, representada na Figura 53, permitiu a identificação de dados considerados irrelevantes, como por exemplo “bom dia”, “obrigada”, “olá”, entre outros. Com esses dados classificados como irrelevantes, reconfigurou-se a *cleaning database* para que os processamentos seguintes os descartassem. A tarefa de identificar este tipo de dados foi um processo recursivo executado por diversas vezes. Neste processo foi essencial a interação humana para analisar sistematicamente os resultados obtidos e verificar a necessidade ou não de algum ajuste. Em caso afirmativo, antes de avançar para a próxima etapa, refazia-se tudo de novo. Neste trabalho de investigação não foi possível, por limitação de espaço, representar todas as redes criadas e analisadas até os dados estarem totalmente refinados. Por esse motivo apresentam-se apenas duas das redes criadas durante todo o processo, a Figura 52 e a Figura 53.

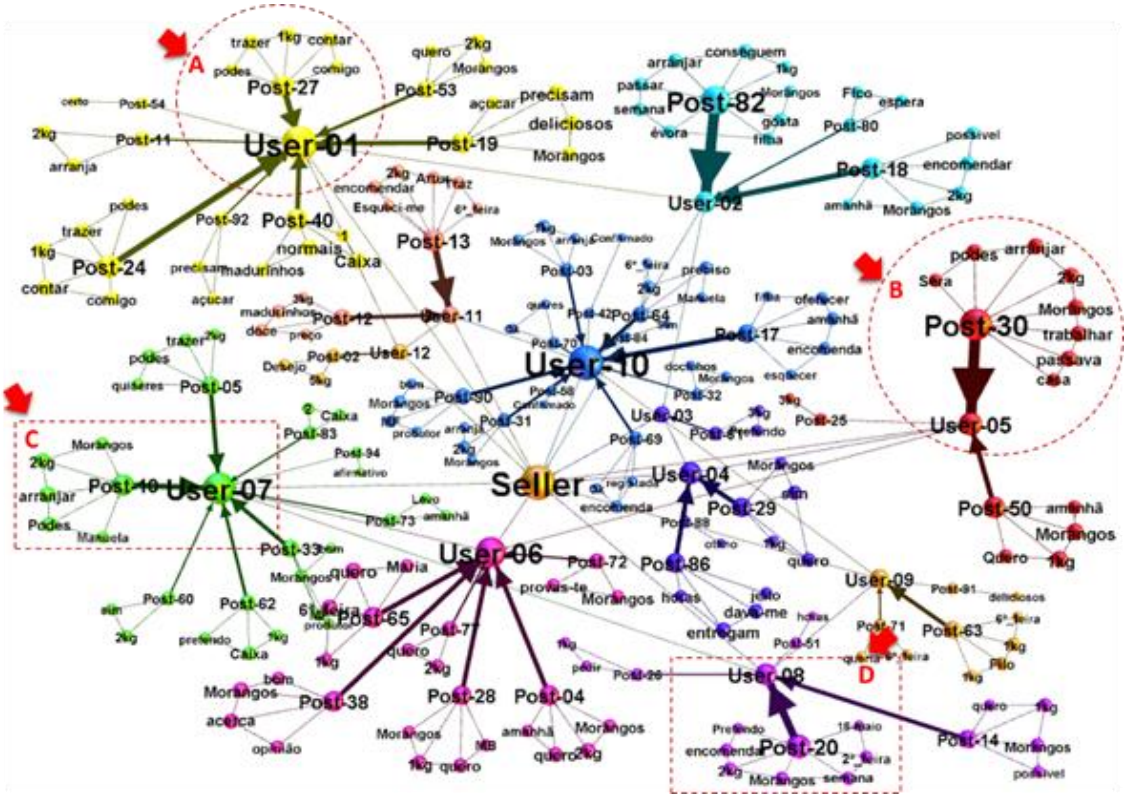
Figura 53 - Rede 6: visualização de dados para *text-mining*.



Por último, a Figura 54 ilustra a criação da rede 7 com todos os utilizadores, *posts* e conceitos, as interações entre os clientes e o vendedor (designado na figura como *Seller*), são mais fáceis de interpretar através da visualização gráfica. Com apoio da SNA,

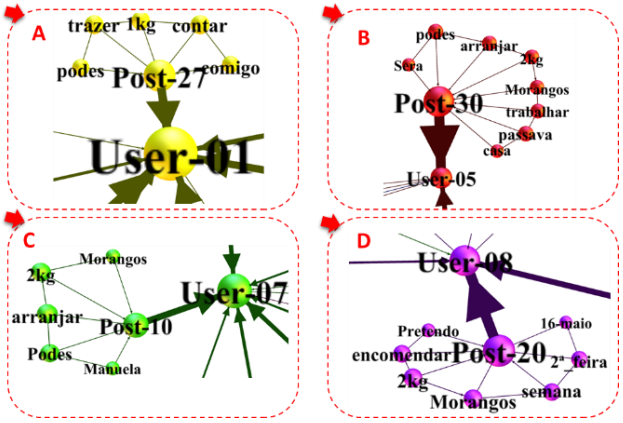
estruturou-se o discurso *web* e os dados encadearam-se de forma coerente para facilitar o seu acesso e análise. Através da visualização do grafo foi possível identificar o relacionamento dos intervenientes, como contribuiram com informação, e o resumo do conteúdo dos *posts*.

Figura 54 - Rede 7: resumo dos pedidos dos clientes



A Figura 55 mostra quatro exemplos de sub-redes semânticas (A, B, C e D) retiradas da rede 7, que resumem algumas das encomendas. Esta rede final era composta por 237 nós e 342 ligações. A sua representação gráfica ofereceu uma visão geral e rápida do conteúdo das trocas discursivas produzidas. O resumo das encomendas auxiliou na identificação das quantidades encomendadas e permitiu identificar quem encomendou o quê, sem perda de informação. Por exemplo, como ilustra a figura, o cliente *User-01*, através da mensagem *post-27* encomendou 1kg de morangos e o cliente *User-08*, através da mensagem *post-20*, encomendou 2kg.

Figura 55 - Exemplo de 4 *posts* resumo de encomendas



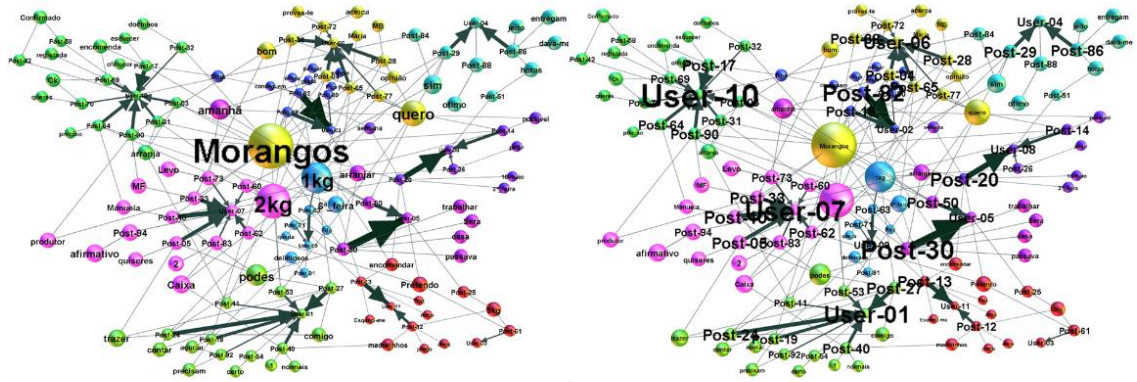
6.3.3 Processamento semântico dos dados

No sentido de atingir o objetivo do problema de decisão definido neste estudo de caso, (estimar as quantidades que cada cliente encomendou), realizou-se uma análise semântica dos *posts* com recurso *ao text-mining* e à SNA. Na análise semântica, e para calcular as quantidades encomendadas, não foram considerados os *posts* do vendedor (administrador do grupo representado como *seller*) e analisaram-se apenas os dos clientes. Após o processamento dos dados, com recurso ao *Excel*, visualizou-se o discurso *web* num grafo para interpretação e análise.

Partindo da premissa de que os processos de extração e análise individual de conceitos forneciam resumos das opiniões expressa no discurso *web*, criaram-se redes de palavras-chave. Pretendeu-se extrair informação útil publicada pelos clientes para identificar as suas encomendas e identificar as quantidades que pretendiam. A representação gráfica da rede 8, na Figura 56, que foi construída com unicidade de conceitos, agregou as três entidades do discurso *web*. Ou seja, dentro da rede semântica global cada conceito foi representado por um nó com um *ID* único, para identificar as quantidades mais encomendadas e os clientes que efetuaram mais encomendas. A Figura 56 ilustra as duas perspetivas da mesma rede, após aplicação das métricas *out-degree* e *in-degree*.

Na rede representada em (a) da Figura 56, a dimensão dos nós e das suas legendas definiu-se com a métrica *out-degree* para identificar as quantidades mais encomendadas. Já na rede (b), representada na mesma figura, dimensionou-se a legenda dos nós com a métrica *in-degree* para identificar os clientes que mais encomendas efetuaram. Como se pode observar, em cada *layout* o tamanho dos nós alterou-se consoante as características específicas da métrica representada graficamente.

Figura 56 - Rede 8: visualização semântica de palavras-chave



(a) Quantidades mais solicitadas (b) Clientes que efetuaram mais pedidos

Para calcular as quantidades encomendadas utilizou-se o *out-degree* pois apenas interessava o nó com o maior número de votos diretos e/ou que podia chegar diretamente a um maior número de outros nós. Para além disso, consideraram-se as variáveis k_i (i=1,…,n) como sendo as quantidades associadas a cada um dos i conceitos. Assim, as quantidades calcularam-se através do produto entre a variável k_i e a métrica *out-degree*, obtendo-se os resultados da Tabela 17 (a). Considere-se, por exemplo, o conceito 2Kg (k_i=2) que registou um valor elevado de 13 na métrica de centralidade *out-degree* (quantidade de morangos encomendados). O resultado obtido, 2×13=26, significou que os clientes efetuaram 13 encomendas em caixas de 2kg o que resultou em 26kg de morangos vendidos.

Para calcular as quantidades que cada cliente encomendou, adicionou-se na *graph database* um atributo que identificou a relação entre as três entidades: *user, post* e *concept*. Esta relação permitiu calcular o *out-degree* da entidade *user* para a entidade *concept* e assim identificar quem escreveu o *concept* e em que *post*. As quantidades por cliente são o produto entre as ocorrências para cada utilizador do conceito e a variável k_i (i=1,…,n), apresentando-se os resultados na Tabela 18 (a). Considere-se, por exemplo, o cliente *User-01*. Este cliente encomendou 2 caixas de 1kg e 2 caixas de 2kg de morangos. Ou seja, a quantidade de morangos que encomendou foi de 6kg.

6.3.4 Comparação dos resultados estimados com os reais

Os dados foram analisados manualmente e a Tabela 17 (b) e a Tabela 18 (b) exibem as quantidades de morangos efetivamente encomendadas. Comparando estas quantidades com as estimativas obtidas utilizando a SNA, constatou-se uma diferença de 2kg.

Tabela 17 - Quantidades de morangos solicitadas por tipo de embalagem

Conceito (k)	(a) Resultados modelo		(b) Resultados manuais	Δ
	Out-degree	Kg	Kg	(a) - (b)
1kg	12	12	10	**2**
2kg	13	26	26	**0**
3kg	3	9	9	**0**
5kg	1	5	5	**0**
Total	**29**	**52**	**50**	**2**

De acordo com a SNA obtiveram-se 52kg de morangos a encomendar, quando na realidade só foram necessários 50kg. A diferença de 2kg resultou da alteração de quantidades encomendadas por parte de clientes. Os utilizadores *User-02* e *User-07* encomendaram menos 1kg cada um do que aquilo que comentaram no *Facebook*.

Tabela 18 - Quantidades de morangos solicitadas pelo cliente

Utilizador	(a) Resultados calculados com métricas da SNA					(b) Pedidos efetivamente emitidos					Δ
	1kg	2kg	3kg	5kg	Total	1Kg	2Kg	3Kg	5Kg	Total	(a) - (b)
User-01	2	4	0	0	**6**	2	4	0	0	**6**	**0**
User-02	1	2	0	0	**3**	0	2	0	0	**2**	**1**
User-03	0	0	3	0	**3**	0	0	3	0	**3**	**0**
User-04	1	0	0	0	**1**	1	0	0	0	**1**	**0**
User-05	1	2	3	0	**6**	1	2	3	0	**6**	**0**
User-06	2	4	0	0	**6**	2	4	0	0	**6**	**0**
User-07	1	6	0	0	**7**	0	6	0	0	**6**	**1**
User-08	2	2	0	0	**4**	2	2	0	0	**4**	**0**
User-09	1	0	0	0	**1**	1	0	0	0	**1**	**0**
User-10	1	4	0	0	**5**	1	4	0	0	**5**	**0**
User-11	0	2	3	0	**5**	0	2	3	0	**5**	**0**
User-12	0	0	0	5	**5**	0	0	0	5	**5**	**0**
Total	**12**	**26**	**9**	**5**	**52**	**10**	**26**	**9**	**5**	**50**	**2**

6.3.5 Conclusões do estudo de caso

Neste estudo de caso utilizou-se o modelo apresentado no capítulo 5. Os dados foram recolhidos, processados e analisados com recurso a ferramentas de SNA (*Netvizz* e *Gephi*). Exploraram-se as redes de duas perspetivas diferentes. Para além da análise das interações sociais entre utilizadores, combinou-se a análise semântica do discurso produzido pelos mesmos. Através desta análise e dos resultados obtidos, concluiu-se que era possível extrair informação para apoio à decisão. Por exemplo, os conceitos 1kg, 2kg, etc., permitiram calcular quantidades a entregar, sendo informação semântica útil para decidir sobre o tipo de caixa a utilizar.

Relativamente às interações sociais, concluiu-se que a SNA traz contribuições para perceber os clientes, pois permite criar novas visões apoiadas em dados reais, fruto das trocas discursivas entre utilizadores. Através das métricas da SNA não só foi possível identificar a estrutura da rede de utilizadores, mas também "traduzir" a rede numa representação gráfica, que pôde ser utilizada para análise. Desta forma, foi possível identificar os utilizadores que atuavam como líderes (que mais encomendas faziam) e/ou que influenciavam outros (recomendação do produto), bem como determinar a força relativa do líder através de um conceito-chave.

No que se refere à análise semântica, para interligar a entidade *concepts* aos *posts* seguiram-se duas abordagens. Com a primeira, obteve-se um resumo de cada troca discursiva e com a segunda identificaram-se as palavras-chave da rede. Exploraram-se os dados, recorrendo ao *text-mining* para construir redes semânticas que resumissem as trocas discursivas entre utilizadores. As técnicas de PLN permitiram converter um *post* em múltiplos fragmentos (tantos quantos os conceitos que existiam num *post*).

Para além disso, o software de SNA que incorpora diversas técnicas e algoritmos de visualização, tais como o *Force Atlas* e o *Fruchterman and Reingold*, ajudaram a "traduzir" os conceitos extraídos dos *posts* em algo mais compacto e, portanto, mais compreensível. Assim, através da SNA, relacionou-se a métrica *out-degree* e palavras-chave (1Kg, 2kg, etc.), concluindo-se que era possível, através da análise semântica, recomendar as quantidades necessárias a transportar e a entregar aos clientes. Identificaram-se os clientes que efetuaram mais encomendas e a margem de erro encontrada, 2kg, não foi considerada significativa.

6.4 Estudo de caso - Volkswagen

No estudo de caso Volkswagen (VW), o foco da análise dos dados foi perceber o impacto do escândalo de emissões de poluentes da VW (em setembro 2015) na *FanPage* da *VolkswagenUK*[68]. Na sequência do escândalo, a VW despediu quadros de topo por presumível envolvimento na instalação de *software* que adulterava as medições de consumo e de emissões poluentes nos veículos, o que fez disparar os custos do grupo por vários motivos (multas, reparações, retomas, etc.).

[68] Acesso em setembro 2015, disponível na *FanPage* da *VolkswagenUK*: https://www.facebook.com/VolkswagenUK/.

Atualmente, opiniões e informações negativas sobre as marcas disseminam-se muito rapidamente nas RSO e a VW não foi exceção. Neste estudo de caso utilizaram-se métricas e técnicas de visualização da SNA para ilustrar o impacto desse escândalo. O objetivo principal foi estruturar os *posts*, produzidos na *FanPage* acima referida, para perceber a reação dos utilizadores, clientes ou não da marca, bem como desenvolver indicadores para monitorizar as conversas *online*. Este tipo de "escuta" das RSO da *web* social permite que as marcas identifiquem informação útil, com o objetivo de limitar danos de imagem e evitar que um acontecimento evolua para uma crise e/ou desastre. Permite também que as empresas estejam sempre informadas acerca do que as pessoas dizem sobre elas, em cada RSO, permitindo assim ações direcionadas em situações críticas.

6.4.1 Extração de dados

Os dados utilizados neste estudo de caso recolheram-se ao longo de 22 semanas com recurso ao *Netvizz* que, de forma simples, permitiu importar os dados de uma página institucional da VW. O assunto foi conhecido a 18 de setembro de 2015 e os dados foram recolhidos em períodos temporais de uma semana, entre 31 de agosto de 2015 e 31 de janeiro de 2016. Os mesmos permitiram interligar várias fontes de dados e construir várias redes e, com isso, descobrir os temas mais discutidos logo após o escândalo.

Na Tabela 19 caracterizam-se as recolhas de dados efetuadas. A primeira coluna identifica o intervalo de tempo da recolha dos dados, a segunda e terceira coluna identificam o início e o fim da recolha, respetivamente. As colunas quatro, cinco e seis indicam o total de *users* e *posts* por *snapshot*. A última coluna indica a quantidade de ligações entre as entidades *user* e *post*. Quando comparamos os vários *snapshots*, observa-se que o terceiro, o oitavo e o décimo quarto, registaram grandes volumes de atividade e fluxos de informação. Nessas semanas o número de utilizadores e mensagens são elevados. Observa-se, ainda, que o número de ligações também é considerável.

Extraíram-se duas fontes de dados com a atividade comunicativa dos utilizadores em torno dos *posts* da *FanPage*. Os dois *outputs* recolhidos continham dados de dois tipos: estruturados e semiestruturados. Os dados estruturados, ficheiro grafo *two-mode* no formato GDF com as entidades *user* (anonimizados) e *post* e as ligações entre as duas. Neste ficheiro, um utilizador encontrava-se ligado a uma mensagem, caso a tivesse comentado ou colocado um *like*. Já o segundo ficheiro, com dados semiestruturados em formato tabular (TSV), continha o texto dos comentários dos utilizadores.

Tabela 19 - Caraterização semanal dos dados recolhidos (nós e ligações)

Semana	Intervalo tempo		Nós			Ligações
	Início	Fim	*user**	*post*		*user* | *post**
				iniciais*	respostas**	
Sem-01	31/ago/15	06/set/15	1.319	44	321	1.370
Sem-02	07/set/15	13/set/15	1.500	28	132	1.546
Sem-03	14/set/15	20/set/15	11.973	36	2.177	15.212
Sem-04	21/set/15	27/set/15	578	105	638	688
Sem-05	28/set/15	04/out/15	383	68	400	466
Sem-06	05/out/15	11/out/15	168	43	173	216
Sem-07	12/out/15	18/out/15	58	53	79	106
Sem-08	19/out/15	25/out/15	3.769	30	1.482	3.789
Sem-09	26/out/15	01/nov/15	809	14	563	819
Sem-10	02/nov/15	08/nov/15	635	14	632	644
Sem-11	09/nov/15	15/nov/15	47	24	50	67
Sem-12	16/nov/15	22/nov/15	20	19	66	36
Sem-13	23/nov/15	29/nov/15	573	19	188	599
Sem-14	30/nov/15	06/dez/15	6.807	19	1.707	6.981
Sem-15	07/dez/15	13/dez/15	431	15	405	479
Sem-16	14/dez/15	20/dez/15	311	17	191	341
Sem-17	21/dez/15	27/dez/15	208	11	119	227
Sem-18	28/dez/15	03/jan/16	1.510	12	259	1.527
Sem-19	04/jan/16	10/jan/16	314	18	113	379
Sem-20	11/jan/16	17/jan/16	249	17	166	289
Sem-21	18/jan/16	24/jan/16	80	23	61	103
Sem-22	25/jan/16	31/jan/16	93	22	69	119

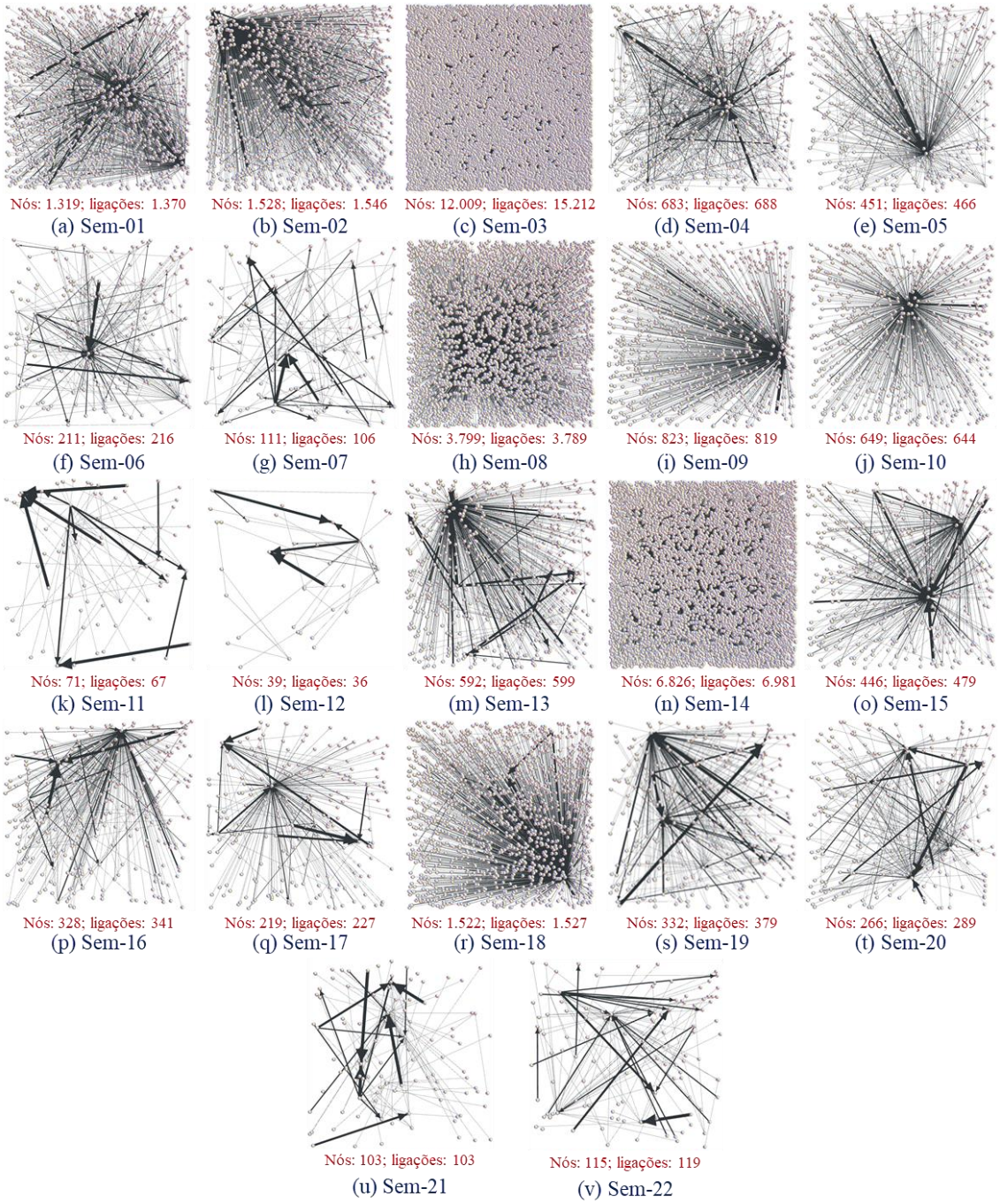
* Dados formato GDF; ** Dados formato TSV.

Para a análise semântica utilizaram-se os conteúdos dos *posts* e respetivas respostas, extraídos como texto puro dos *output*s em GDF e TSV. Assim, garantiu-se que só entravam os recursos linguísticos textuais da comunicação na análise. Estes dados lexicais eram dados não estruturados, o que implicou o seu processamento, limpeza e estruturação.

6.4.2 Processamento e interpretação dos dados

Num primeiro passo, à medida que as recolhas eram efetuadas, os dados originais em formato GDF, e sem qualquer processamento, foram importados para o *Gephi* para visualização. As representações gráficas dos vários *snapshots* da Figura 57, auxiliaram na interpretação semanal das redes. Em cada *layout* indicou-se a quantidade de nós e ligações entre eles.

Figura 57 - Visualização dos dados não processados



Apesar da reduzida dimensão das representações gráficas da Figura 57, são visíveis algumas características das redes aí apresentadas. Numa primeira análise à figura, observa-se que existem grandes diferenças entre os *layouts* que representam as 22 semanas analisadas. Considere-se, por exemplo, a rede (c) da Figura 57 que representa a atividade produzida na semana do acontecimento. Comparativamente com as duas semanas anteriores, na semana do escândalo (*snapshot* Sem-03), o número de novos utilizadores na *FanPage* da VW aumentou consideravelmente. Passou de 1.500 utilizadores no *snapshot* da semana anterior,

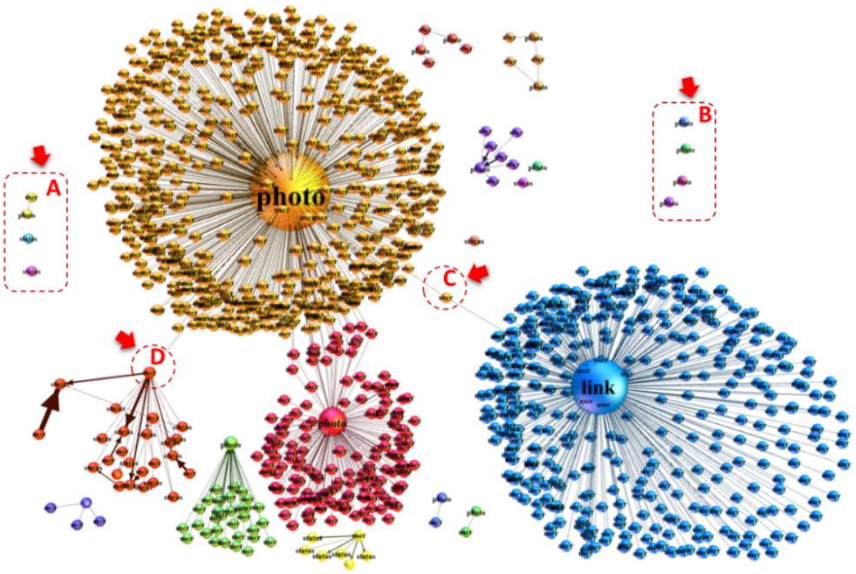
para 11.973. Para além disso, o fluxo de informação entre partilhas de *links* e mensagens criadas e comentadas cresceu proporcionalmente, porque as pessoas recorriam à RSO para manifestar a sua insatisfação. As *ligações* passaram de 1.546 para 15.212 e os *posts* de 160 para 2.212.

A visualização dos *layouts* anteriores permitiu ainda outras interpretações. Verificou-se que todas as redes tinham topologias e estruturas distintas. Observou-se que os grafos apresentados em (c), (h) e (n), respetivamente, eram redes muito densas. Estes grafos, que evidenciavam estruturas de rede densas compostas pelas ligações existentes entre os nós, indicavam a capacidade que os utilizadores tinham em estabelecer ligações dentro e fora da *FanPage*. Em sentido oposto, verificou-se que os grafo (l) e (k) eram menos densos o que permitiu de imediato identificar o sentido e a força das ligações entre os nós (espessura das setas).

A utilização do *Gephi* permitiu não só produzir representações gráficas das redes para identificar subconjuntos específicos de nós, através da atribuição de cores e legendas, mas também identificar *posts* irrelevantes e utilizadores-chave. Do total de 10.642 mensagens identificaram-se e retiraram-se da análise 1.337 *posts* irrelevantes. Estes *posts* caracterizavam-se por serem fotos, vídeos, ou não incluírem nenhuns recursos linguísticos textuais.

Como se pode observar pelo grafo do *snapshot* Sem-01, apresentado na Figura 58, identificaram-se vários grupos distintos (laranja, verde, azul, etc.). Visualmente conseguiu associar-se a entidade *post* à entidade *user*, o que simplificou o processo de captar e interpretar a informação com base nas características dos dados. Nesta rede e de acordo com as técnicas de representação gráfica da SNA, observou-se que os utilizadores do exemplo C e D eram utilizadores-chave visto que, se as suas ligações deixassem de existir, a rede dividia-se em múltiplas sub-redes isoladas. Para além disso, observou-se que existiam *posts* irrelevantes como por exemplos os assinalados em A e B.

Figura 58 - Visualização gráfica para auxílio ao *data-mining*



Nesta etapa de processamento, após a interpretação dos dados originais e o seu armazenamento, criaram-se atributos para interligar os vários conjuntos de dados. Em *Excel*, a base de dados estruturou-se num único ficheiro composto por várias folhas de cálculo que armazenaram os dados de acordo com a sua função. A Tabela 20 mostra a estrutura da base de dados e as folhas de cálculo para armazenamento e processamento das três entidades do discurso *web* deste estudo de caso. A estrutura criada em *Excel* para a primeira semana foi depois replicada, para todos os *snapshots*, garantindo assim que os dados eram agregados, processados e analisados de acordo com as mesmas regras e critérios.

Tabela 20 - Estrutura do suporte de armazenamento dos dados

#	Folhas de cálculo	Descrição
1	Configurações	Variáveis pré-definidas e utilizadas para cálculos.
2	Tabelas Entidades	Tabelas com as entidades dos discurso *web*: *user*, *post* e *concept*.
3	Dados nós	Entrada de dados, processamento e transformação do modelo relacional para *graph database*.
4	Dados comentários	
5	Dados conceitos	
6	Dados ligações sem unicidade	Dados em formato grafo para criação de redes resumo.
7	Dados ligações com unicidade	Dados em formato grafo para criação de redes de palavras-chave.
8	*Cleaning database*	Tabelas de: *smiles*, *stopwords*, pontuação e sinónimos.

O processamento dos dados envolveu a criação de tabelas para armazenar a informação dos utilizadores, o conteúdo das mensagens (*post* e *comment*) e os conceitos. Em cada uma delas criaram-se *IDs* distintos que permitiram não só identificar cada entidade do discurso *web*, mas também validar a sua existência ou não em semanas anteriores. A estrutura dos *IDs* iniciava com uma letra, U para *user*, P para *post*, R para *comment* e C para

concept, seguindo-se uma parte numérica que variava no intervalo [1-99999]. Por exemplo, U-00001 para a entidade *user* e C-00001 para a entidade *concept*. Relativamente à entidade *post* fez-se a distinção entre a mensagem inicial (exemplo P-00001) e os seus comentários (exemplo R-00001) porque, apesar de ambos serem mensagens, encontravam-se em fontes de dados distintas.

A transformação dos dados do modelo relacional para o modelo *graph database*, relacionamentos do tipo $n \times n$, executou-se com recurso a atributos definidos para o efeito. As únicas ligações já estruturadas e existentes nos dados originais eram as ligações entre a entidade *user* e o *post* inicial. Como definido no modelo (subcapítulo 5.2.1 e subcapítulo 5.2.2), as ligações em falta que permitiriam estruturar o discurso inferiram-se através dos *IDs* das entidades *user*, *post* e *concept*. O resultado final, após a transformação dos dados, foi um *output* de dados com todos os nós das três entidades e outro com todas as ligações. Isto é, dados estruturados em formato grafo.

Somente depois da estruturação de todos os dados, oriundos quer das fontes de dados estruturadas (utilizadores e trocas discursivas), quer do processamento semântico (texto contido nos *posts*), foi possível construir os dois tipos de redes semânticas: redes com o resumo das trocas discursivas e redes com palavras-chave (como representado nas Figura 59 e Figura 60, respetivamente).

Na Tabela 21 caracterizaram-se o tipo e número de nós obtidos, após o processamento final dos dados, em cada *snapshot*. A coluna (b) indica o total de utilizadores. As colunas (c) e (d) discriminam os elementos da entidade *post*, por *post* inicial e respostas obtidas respetivamente. As colunas (e) e (f) indicam o total de conceitos para a criação das redes semânticas. Por último, a coluna (g) indica o total de nós para construção da rede semântica de palavras-chave e a coluna (h) para a construção da rede semântica de resumos.

A Tabela 22 foi usada para caracterizar o tipo e a quantidade de ligações obtidas, após o processamento final dos dados e para cada *snapshot*. As colunas (b), (d) e (f) discriminam as ligações entre os elementos intra-entidade (*user*, *post* e *concept*), ou seja, ligações definidas como *one-mode*. Nas colunas (c) e (e) caracterizam-se as ligações entre entidades *user* e *post*, *post* e *concept*, ou seja, ligações inter-entidade definidas como *two-mode*. De referir que as ligações *one-mode* da entidade *concept* (*concept* | *concept*), apresentadas na coluna (g), não foram utilizadas na construção da rede semântica de palavras-chave. Essas ligações, coluna (h), só foram necessárias para a criação das redes

semânticas de resumos, pois o objetivo foi identificar os conceitos imediatamente antes e o imediatamente depois.

Tabela 21 - Rede resumo: tipo e número de nós em cada semana

Nós por entidade

(a) Semana	(b) *user*	(c) *post* iniciais	(d) *post* respostas	(e) *concept**	(f) *concept***	(g)=(b)+(c)+(d)+(e) Total*	(h)=(b)+(c)+(d)+(f) Total**
Sem-01	400	12	294	604	1.094	**1.310**	**1.800**
Sem-02	1.479	14	130	773	1.842	**2.396**	**3.465**
Sem-03	11.876	23	1.758	2.462	8.810	**16.119**	**22.467**
Sem-04	493	33	607	2.363	8.079	**3.496**	**9.212**
Sem-05	355	29	361	1.614	4.605	**2.359**	**5.350**
Sem-06	129	15	166	1.271	2.963	**1.581**	**3.273**
Sem-07	47	15	70	620	1.099	**752**	**1.231**
Sem-08	3.746	12	1.434	3.514	16.879	**8.706**	**22.071**
Sem-09	807	5	540	1.769	6.326	**3.121**	**7.678**
Sem-10	633	5	621	2.204	7.890	**3.463**	**9.149**
Sem-11	25	12	46	417	763	**500**	**846**
Sem-12	16	8	65	590	1.201	**679**	**1.290**
Sem-13	517	10	174	921	1.979	**1.622**	**2.680**
Sem-14	6.807	13	1.424	2.548	7.820	**10.792**	**16.064**
Sem-15	429	7	395	1.650	4.652	**2.481**	**5.483**
Sem-16	311	12	188	880	1.885	**1.391**	**2.396**
Sem-17	208	9	117	794	1.747	**1.128**	**2.081**
Sem-18	1.502	7	250	1.091	2.619	**2.850**	**4.378**
Sem-19	314	10	99	658	1.188	**1.081**	**1.611**
Sem-20	184	10	158	1.195	2.638	**1.547**	**2.990**
Sem-21	60	11	60	565	981	**696**	**1.112**
Sem-22	92	13	63	586	1.149	**754**	**1.317**

*Rede de palavras-chave; **Rede resumo do discurso.

Tabela 22 - Rede resumo: tipo e número de ligações semanais

Ligações entre entidades

(a) Semana	(b) *user* | *user*	(c) *user* | *post*	(d) *post* | *post*	(e) *concept* | *post*	(f) *concept* | *concept*	(g)=(b)+(c)+(d)+(e) Total*	(h)=(b)+(c)+(d)+(f) Total**
Sem-01	255	486	51	1.023	902	**1.815**	**2.717**
Sem-02	70	1.653	128	1.521	1.703	**3.372**	**5.075**
Sem-03	1.601	16.867	1.751	8.271	7.139	**28.490**	**35.629**
Sem-04	316	1.141	581	7.318	7.450	**9.356**	**16.806**
Sem-05	221	765	349	4.156	4.222	**5.491**	**9.713**
Sem-06	103	314	154	2.587	2.782	**3.158**	**5.940**
Sem-07	36	126	61	1.007	1.019	**1.230**	**2.249**
Sem-08	1.143	5.186	1.429	15.424	15.450	**23.182**	**38.632**
Sem-09	300	1.347	535	5.695	5.789	**7.877**	**13.666**
Sem-10	302	1.259	617	7.188	7.276	**9.366**	**16.642**
Sem-11	28	84	42	690	707	**844**	**1.551**
Sem-12	19	88	58	1.028	1.128	**1.193**	**2.321**
Sem-13	134	712	170	1.775	1.798	**2.791**	**4.589**
Sem-14	1.449	8.400	1.417	7.344	6.430	**18.610**	**25.040**
Sem-15	194	864	392	4.195	4.259	**5.645**	**9.904**
Sem-16	110	525	186	1.698	1.689	**2.519**	**4.208**
Sem-17	54	344	116	1.546	1.622	**2.060**	**3.682**
Sem-18	193	1.766	246	2.351	2.364	**4.556**	**6.920**
Sem-19	71	466	88	1.100	1.084	**1.725**	**2.809**
Sem-20	75	379	153	2.322	2.473	**2.929**	**5.402**
Sem-21	34	138	55	896	912	**1.123**	**2.035**
Sem-22	46	176	58	971	1.074	**1.251**	**2.325**

*Rede de palavras-chave; **Rede resumo do discurso.

6.4.3 Processamento semântico

O processamento semântico e a consequente análise dos *posts* publicados *online* na VW, teve como objetivo ilustrar o impacto do escândalo de emissões de poluentes. Para o efeito utilizaram-se métricas e técnicas de visualização da SNA para estruturar os conteúdos das mensagens, encadeando-os às outras duas entidades do discurso *web*. Isso permitiu uma análise *end-to-end* do discurso *web*, com as entidades *user*, *post* e *concept*, que evidenciou as relações estruturais mais importantes e a informação existente dentro das redes.

Para extração dos conceitos dos conteúdos dos *posts* utilizou-se a *cleaning database* e os algoritmos criados para o efeito. A estrutura da *cleaning database* alimentou-se com listas de padrões predefinidos noutros sistemas e com os conceitos extraídos dos conteúdos dos *posts*, como especificado no modelo. Para além disso, a *cleaning database* utilizou-se para a uniformização e padronização do texto, remoção de espaços e eliminação de conceitos irrelevantes.

As mensagens recolhidas encontravam-se escritas em inglês. Por esse motivo, configurou-se a *cleaning database* e foram utilizadas listas de *stopwords* para esse idioma. Os 22 conjuntos de dados iniciais, *snapshots*, eram compostos por um total de 10.642 mensagens, das quais 651 eram *posts* iniciais e 9.991 eram comentários aos mesmos. Das fontes de dados (ficheiros GDF e TSV) selecionaram-se as mensagens que continham texto puro e criaram-se redes semânticas para resumir as trocas discursivas e identificar palavras-chave. Os conjuntos de dados válidos resultantes dos 22 *snapshots*, compostos por 9.305 mensagens, foram processados com o *Excel*.

Antes da otimização da *cleaning database*, as mensagens dos 22 *snapshots*, continham um total de 256.395 conceitos e, após a eliminação dos dados irrelevantes, restaram 88.209. Descartaram-se 66% dos conceitos por terem sido considerados irrelevantes e, apesar da sua eliminação, não retiraram sentido aos resumos finais obtidos. Os *outputs* obtidos com dados semânticos em formato *graph database*, utilizaram-se para construir as redes do discurso *web*.

Assim, como nos estudos de caso anteriores, a abordagem utilizada para agregação dos três níveis do discurso *web* foi a transformação de redes *two-mode* numa única rede *one-mode*. Os resultados finais foram redes com as três entidades em análise (*user*, *post* e *concept*). A criação destas redes envolveu a agregação dos dados dos utilizadores, *posts* e conteúdo textual dos *posts*.

6.4.4 Redes semânticas do discurso *web*

Após o armazenamento e o processamento final, os dados foram explorados e interpretados no *Gephi* com o objetivo de visualizar e manipular de forma simples as várias redes e aplicar técnicas de representação da SNA. Para cada *snapshot*, as métricas da SNA permitiram identificar os nós mais relevantes de cada entidade, classificá-los de acordo com a sua relevância, mostrar e descrever os resultados observados. O *Gephi* permitiu aplicar automaticamente classificações numéricas para a aparência visual das redes em termos de cores e tamanho das entidades do discurso *web*.

Utilizou-se a SNA para identificar os utilizadores mais relevantes, ao nível da sua posição estrutural na rede e ao nível dos conteúdos dos seus *posts*. A Tabela 23 resume a presença dos 15 utilizadores mais assíduos na RSO durante as 22 semanas. Pela análise da tabela observa-se que alguns indivíduos eram utilizadores muito assíduos da *FanPage*. Por exemplo, em 17 das 22 semanas os utilizadores U-00095 e U-00908 publicaram ou comentaram pelo menos um *post*. Este tipo de indicador é importante pois estes utilizadores têm maior poder e tendem a influenciar os outros positiva ou negativamente.

Tabela 23 - Os 15 utilizadores mais assíduos

Entidade / Semana																								
User	Sem-01	Sem-02	Sem-03	Sem-04	Sem-05	Sem-06	Sem-07	Sem-08	Sem-09	Sem-10	Sem-11	Sem-12	Sem-13	Sem-14	Sem-15	Sem-16	Sem-17	Sem-18	Sem-19	Sem-20	Sem-21	Sem-22	Total
VW	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**22**
U-00908	1	1	1		1			1	1	1			1	1	1	1	1	1	1	1	1	1	**17**
U-00095		1	1	1	1	1		1		1	1	1	1	1		1		1	1	1	1	1	**17**
U-04951			1	1	1			1	1	1				1	1	1	1	1	1	1	1	1	**15**
U-07288			1		1	1		1	1	1			1	1	1	1	1	1	1	1			**14**
U-01116					1			1	1	1			1	1	1		1	1	1	1	1	1	**13**
U-00913	1	1	1	1				1	1	1			1	1				1	1	1	1		**13**
U-00445	1	1	1	1				1	1	1			1	1	1	1		1	1				**13**
U-04733			1					1					1	1	1	1	1	1	1	1	1	1	**12**
U-00356			1	1	1					1			1	1	1	1	1	1	1	1			**12**
U-05925			1	1	1	1		1		1	1			1	1	1						1	**11**
U-04424			1	1	1	1							1	1				1	1	1	1	1	**11**
U-04502			1		1	1		1		1			1	1			1	1			1	1	**11**
U-00025	1	1	1					1					1	1			1		1	1	1	1	**11**
U-09945			1		1			1		1				1	1	1	1	1	1	1			**11**
U-02519		1	1		1	1		1	1	1			1					1		1		1	**11**
U-02685				1		1	1				1			1					1	1	1	1	**11**

As métricas de centralidade do utilizador VW (dono da *FanPage*), resumidas na Tabela 24, permitiram identificar a atividade e o controlo do fluxo de informação deste utilizador. Considere-se, por exemplo, o *snapshot* Sem-03 da Tabela 24, período onde ocorreu o acontecimento. Nessa semana registaram-se os valores mais elevados nas métricas de centralidade *in-degree* (quantidade de *posts* ou comentários publicados na *FanPage*) e *out-degree* (quantidade de respostas dadas pela VW). Já relativamente às restantes métricas,

os valores mais elevados de *betweenness centrality*, *PageRank* e *eigenvector* registaram-se no *snapshot* Sem-01. No que se refere à métrica *closeness centrality*, o valor mais elevado registou-se no *snapshot* Sem-09. Estes indicadores sustentados pela SNA evidenciaram a capacidade de atendimento e resposta dada aos utilizadores (clientes ou não da marca).

Tabela 24 - *User* VW métrica SNA

Semana	*Degree*	*In-degree*	*Out-degree*	*Closeness*	*Betweenness*	*PageRank*	*Eigenvector*
Sem-01	264	222	42	0,711	0,009	0,037	1,000
Sem-02	87	35	52	0,711	0,001	0,002	0,028
Sem-03	1.735	1.563	172	0,799	0,001	0,014	0,277
Sem-04	323	212	111	0,639	0,004	0,007	0,188
Sem-05	215	152	63	0,650	0,003	0,008	0,166
Sem-06	100	66	34	0,800	0,001	0,005	0,144
Sem-07	35	0	35	0,754	0,000	0,001	0,000
Sem-08	1.179	1.117	62	0,851	0,001	0,016	0,208
Sem-09	317	292	25	0,875	0,001	0,011	0,178
Sem-10	346	285	61	0,849	0,002	0,009	0,174
Sem-11	45	0	45	0,778	0,000	0,001	0,000
Sem-12	45	0	45	0,718	0,000	0,001	0,000
Sem-13	148	115	33	0,683	0,002	0,009	0,175
Sem-14	228	182	46	0,869	0,002	0,008	0,208
Sem-15	119	78	41	0,717	0,003	0,007	0,175
Sem-16	75	31	44	0,779	0,001	0,003	0,115
Sem-17	201	173	28	0,765	0,001	0,008	0,101
Sem-18	90	55	35	0,804	0,002	0,007	0,247
Sem-19	107	53	54	0,812	0,001	0,004	0,177
Sem-20	44	5	39	0,760	0,001	0,001	0,026
Sem-21	60	18	42	0,797	0,002	0,004	0,162
Sem-22	1.483	1.424	59	0,738	0,001	0,019	0,200

Relativamente à Tabela 25, evidenciam-se para cada *snapshot* os utilizadores com valores mais elevados nas métricas de centralidade, comparativamente com todos os outros. Pela observação da tabela, verificou-se que a VW só registou valores máximos na métrica *degree*. Nas restantes, estes resultados indicaram que outros utilizadores, que não a VW, tinham mais poder no controlo do fluxo de informação e eram mais ativos.

Tabela 25 - Utilizadores com valor máximo em cada métrica

Semana	*Degree*	*In-degree*	*Out-degree*	*Closeness*	*Betweenness*	*PageRank*	*Eigenvector*
Sem-01	VW	VW	VW	*	VW	VW	VW
Sem-02	VW	VW	VW	*	VW	VW	VW
Sem-03	VW	VW	VW	*	VW	VW	VW
Sem-04	VW	VW	VW	*	VW	VW	VW
Sem-05	VW	VW	VW	U-14204	VW	VW	VW
Sem-06	VW	VW	VW	U-17828	VW	VW	VW
Sem-07	VW	U-14720	VW	*	U-14593	U-14720	U-14720
Sem-08	VW	VW	VW	VW	VW	VW	VW
Sem-09	VW	VW	U-14689	VW	VW	VW	VW
Sem-10	VW	U-17796	VW	U-17796	VW	VW	U-17796
Sem-11	VW	U-18.262	VW	*	U-18262	U-18281	U-18281
Sem-12	VW	U-18293	VW	U-00095	U-00095	U-18293	U-18293
Sem-13	VW	VW	VW	U-12973	VW	VW	VW
Sem-14	VW	VW	VW	*	VW	VW	VW
Sem-15	VW	VW	VW	U-24.465	VW	VW	VW
Sem-16	VW	VW	VW	*	VW	VW	VW
Sem-17	VW	VW	VW	*	VW	VW	VW
Sem-18	VW	VW	VW	*	VW	VW	VW
Sem-19	VW	VW	VW	*	VW	VW	VW
Sem-20	VW	VW	VW	*	VW	VW	VW
Sem-21	VW	U-17796	VW	*	VW	VW	U-17796
Sem-22	VW	VW	VW	*	VW	VW	VW

* Métrica evidenciou vários utilizadores com valor elevado igual a 1

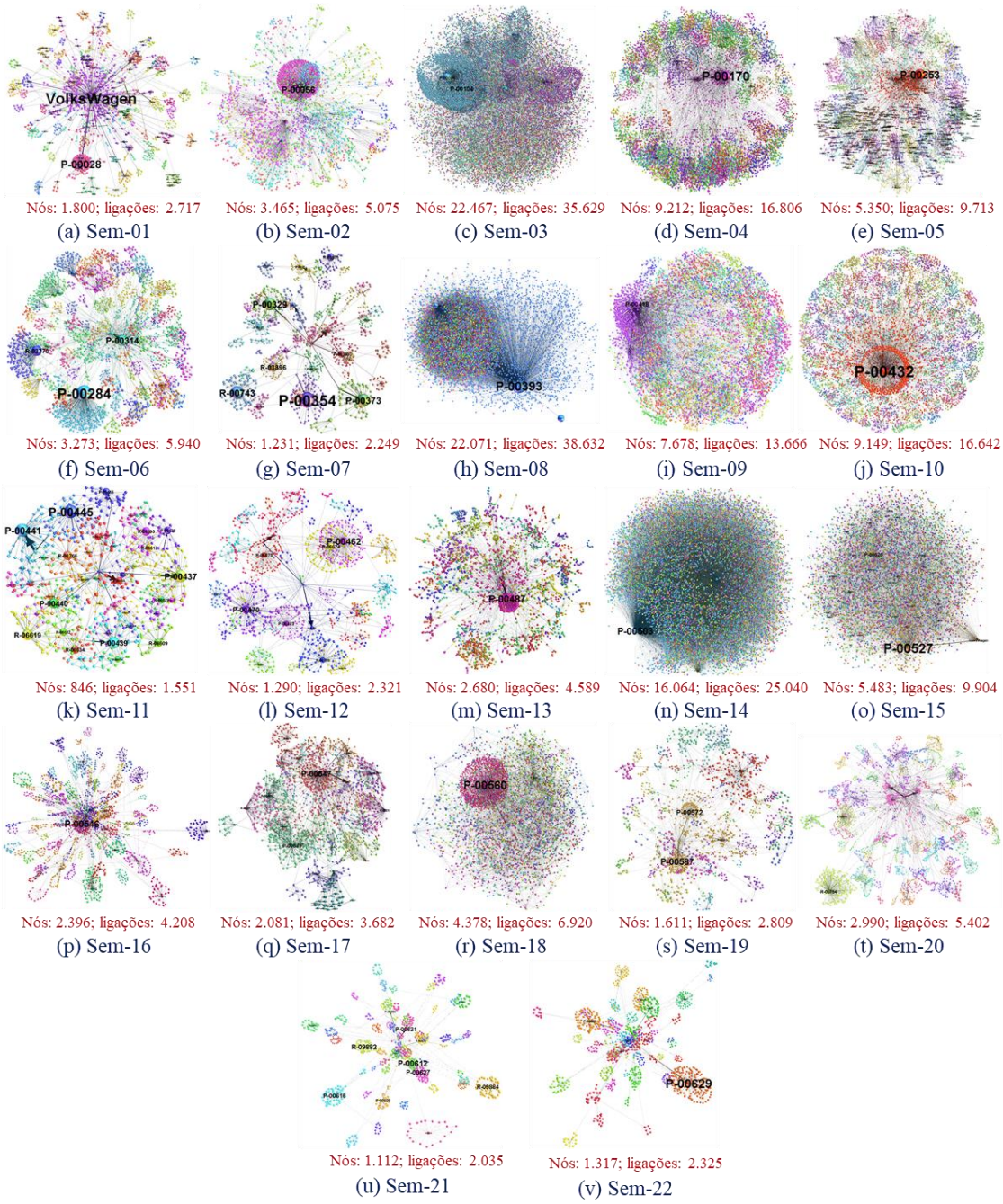
As redes semânticas com o resumo das trocas discursivas e com as palavras-chave, exibidas nas Figura 59 e Figura 60 respetivamente, ilustram as representações gráficas dos 22 cortes transversais da rede após aplicação das métricas de centralidade da SNA. A representação gráfica destes cortes, permitiu mostrar, sequencialmente, resultados relacionados e delimitados pelos intervalos de tempo definidos, tanto para as redes semânticas de resumos como para as de palavras-chave. Também nestas redes, os subgrupos identificaram-se com a métrica *modularity class*.

Como definido no modelo, construíram-se redes semânticas para resumir os conteúdos de cada *post* através de uma sub-rede semântica. A sub-rede semântica construiu-se estabelecendo uma relação entre os pares de *concept* contidos nos *posts*. Como se pode observar nos *layouts* da Figura 59, o tamanho dos nós alterou-se de acordo com as propriedades da rede e das métricas utilizadas. Nestas redes, os *posts* mais comentados ou visualizados identificaram-se com a métrica *in-degree*. Para além disso, o encadeamento das três entidades do discurso *web* numa única rede ofereceu uma visão geral dos interesses, pontos de vista e da posição (a favor ou contra) dos utilizadores da RSO.

Através da análise visual das redes apresentadas na Figura 59 foi possível constatar que três delas eram muito densas e evidenciavam volumes de informação elevados. Os

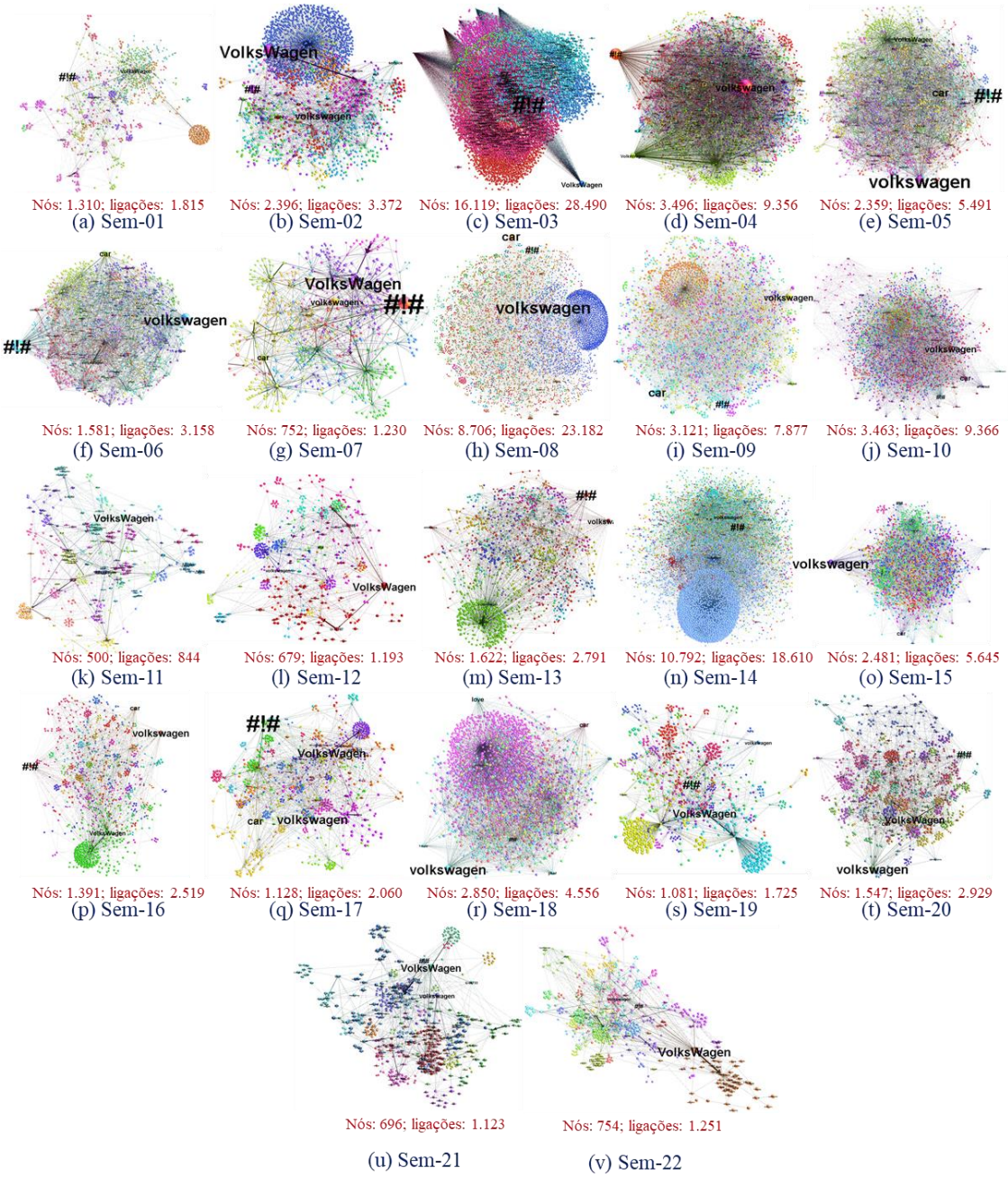
zooms aos *layouts*, utilizando o *Gephi*, facilmente permitiram identificar os diferentes *posts*, e os conceitos neles contidos. A aplicação do algoritmo da *modularity class* agregou todos os conceitos que pertenciam a um mesmo *post*, interpretando-os como subgrupos. A estruturação do discurso *web*, com recurso à SNA, facilitou o encadeamento dos dados e a sua análise.

Figura 59 - Redes resumo do discurso *web*



Enquanto que nas redes da figura anterior cada *post* tinha associada uma sub-rede semântica, na rede de palavras-chave qualquer *post* estava ligado à rede semântica global. Neste tipo de rede, as trocas discursivas entre utilizadores foram estruturadas para criar redes semânticas de palavras-chave. O objetivo foi identificar palavras-chave, tanto para alimentar a *cleaning database*, como para identificar informação útil para apoio à decisão. A extração de palavras-chave foi importante para identificar os conceitos mais utilizados e relevantes nos *posts*.

Figura 60 - Redes de palavras-chave

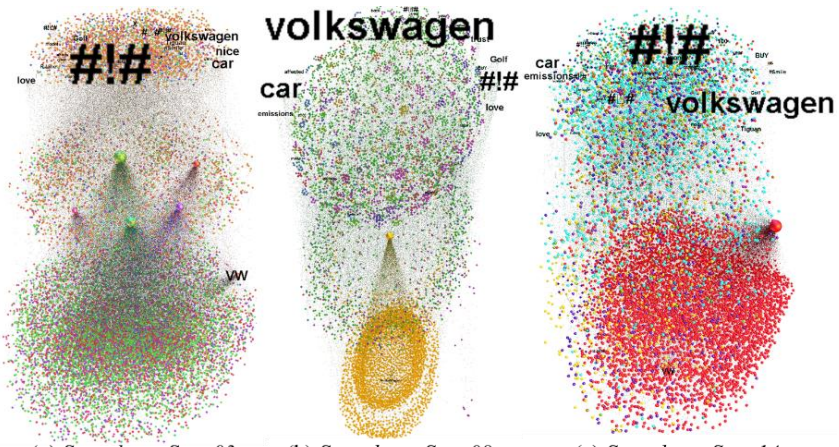


Nas redes representadas na Figura 60, a dimensão dos nós e das suas legendas definiu-se com a métrica *out-degree* para identificar os conceitos mais utilizados. Neste estudo de caso, a análise do conteúdo semântico permitiu a extração de informação útil das mensagens trocadas pelos utilizadores e tirar algumas conclusões sobre o que estava a ser dito. Através de uma análise textual, extraiu-se uma lista das palavras-chave mais utilizadas e a sua classificação.

As redes de palavras-chave da Figura 60, para além de evidenciarem informação semântica relevante, evidenciaram a estrutura e as relações que existiam entre os *posts* e os seus conceitos. Apesar de alguns dos *layouts* apresentarem estruturas muito densas, as métricas da SNA capturaram diferenças importantes, não só para a análise estrutural, mas também para a semântica.

Neste estudo de caso não foi possível, por limitação de espaço[69], representar todas as redes criadas e analisadas. Por esse motivo apresentaram-se apenas na Figura 61 as três redes de palavras-chave dos *snapshots* de maior fluxo e volume de informação, (a) Sem-03, (b) Sem-08 e (c) Sem-14. A sua representação gráfica ofereceu uma visão geral do conteúdo das trocas discursivas e permitiu identificar os conceitos mais utilizados. A análise dos conteúdos semânticos e a identificação de palavras-chave tais como: *emissions*, *problem*, *affected*, permitiu obter informação relevante dos tópicos mais comentados. Como ilustrado na figura, nestes *snapshots* curiosamente os conceitos *love* e *nice* foram identificados como palavras-chave na semana do escândalo.

Figura 61 - Representação gráfica das 3 semanas com maior fluxo de comunicação



(a) *Snapshot – Sem-03* (b) *Snapshot – Sem-08* (c) *Snapshot – Sem-14*

[69] Redes e resultados das métricas da SNA, redes de resumo e de palavras-chave das 22 semanas, disponível em: https://meocloud.pt/link/9e4e90c5-7065-4ec6-9286-b85b7b98c1c9/Extratexto/.

Pela observação dos *layouts* da Figura 61, foi interessante notar que o ponto de exclamação (apesar de não ser um conceito ou palavra) foi muito utilizado. Como referido anteriormente, a utilização de conjuntos de caracteres tais como "!!!!", "????" e/ou "…" tem o objetivo de enfatizar uma ideia e a sua inclusão na análise permite perceber o grau de satisfação e/ou insatisfação dos utilizadores. A elevada utilização, por parte dos utilizadores, do recurso linguístico ponto de exclamação, representado na figura pela sequência #!#, foi para, de forma rápida, expressarem algum desagrado e insatisfação associados à emissão de poluentes. Neste estudo de caso, e como definido no modelo, os conjuntos de caracteres ".", "?" e "!" substituíram-se pelo termo padronizado "#.#", "#!#" e "#?#" respetivamente.

Através da métrica de centralidade *out-degree* foi possível extrair uma lista das palavras-chave mais utilizadas e classificá-las. A Tabela 26 apresenta os 15 conceitos mais utilizados nas trocas discursivas das 22 semanas analisadas.

Tabela 26 - Os 15 conceitos mais utilizados em cada semana

Entidade / Semana																							
Conceito	**Sem-01**	**Sem-02**	**Sem-03**	**Sem-04**	**Sem-05**	**Sem-06**	**Sem-07**	**Sem-08**	**Sem-09**	**Sem-10**	**Sem-11**	**Sem-12**	**Sem-13**	**Sem-14**	**Sem-15**	**Sem-16**	**Sem-17**	**Sem-18**	**Sem-19**	**Sem-20**	**Sem-21**	**Sem-22**	**Total**
#!#	64	44	438	223	117	77	40	423	175	183	21	23	75	258	119	61	52	67	40	46	25	30	**2.601**
VolksWagen	19	48	166	241	123	81	23	588	164	224	19	26	56	168	135	61	43	96	25	53	26	24	**2.409**
car	14	25	176	157	100	47	22	451	188	187	10	11	22	109	111	46	39	60	15	35	11	17	**1.853**
#...#	16	13	129	74	34	15	7	179	66	75	4	-	15	-	42	19	11	16	7	14	7	7	**750**
Golf	5	6	97	62	25	7	-	189	32	49	3	1	9	29	43	8	9	38	4	9	10	7	**642**
emissions	-	1	14	73	41	20	5	155	65	41	1	2	29	72	45	6	11	16	6	15	5	-	**623**
love	-	2	128	38	20	5	-	184	20	32	1	-	8	53	16	3	1	56	6	1	-	1	**575**
#!!#	3	6	96	29	22	9	3	68	37	32	3	2	7	56	20	6	7	6	4	5	3	3	**427**
back	4	8	31	34	28	12	8	90	30	38	2	8	6	26	22	14	11	11	6	10	3	7	**409**
affected	-	-	-	31	28	9	4	113	80	106	1	1	3	8	11	2	4	4	1	3	-	-	**409**
problem	6	14	22	50	26	9	4	70	34	43	5	3	6	25	24	10	5	16	5	14	7	7	**405**
us	4	15	31	49	18	12	10	44	17	43	5	13	6	34	18	8	11	9	9	17	10	9	**392**
#!!!#	3	9	90	19	29	10	-	74	31	24	-	1	8	51	13	4	2	7	3	7	-	2	**387**
BUY	2	5	17	33	23	9	-	94	34	28	-	4	7	48	32	8	7	5	5	7	-	1	**369**
time	11	9	32	32	15	11	3	70	19	39	5	10	16	24	14	9	8	7	5	9	7	3	**358**
Year	3	3	31	29	8	7	3	94	15	21	2	1	10	35	18	5	8	45	4	13	1	2	**358**

6.4.5 Conclusões do estudo de caso

A reputação da marca VW, assente na confiança dos clientes, ficou exposta e afetada devido ao escândalo das emissões. A análise dos conteúdos semânticos dos *posts* e a identificação de palavras-chave permitiu obter informação relevante dos tópicos em discussão e perceber que quem entrava na *FanPage* da VW à procura de uma qualquer outra informação sobre a marca ficava inevitavelmente exposto à história do escândalo.

Curiosamente, verificou-se que os conceitos *love* e *nice* foram muito utilizados. Se por um lado a SNA evidenciou conceitos mais negativos, tais como: *emissions*, *problem*, *affected*, por outro também fez sobressair informação positiva. Uma interpretação possível foi que independentemente de os utilizadores estarem preocupados com as emissões de gases poluentes, a reputação da marca VW (na Europa) não ficou muito afetada apesar de ter ficado exposta. Tanto assim é que, de acordo com a comunicação social *online*[70], as vendas da VW, na União Europeia têm aumentado.

Apesar deste tipo de situações serem esporádicas, as suas consequências podem perpetuar-se no tempo (como foi o caso). Por esse motivo, há necessidade de uma monitorização contante das RSO no sentido de melhorar e apoiar a tomada de decisão para limitar eventuais danos. Para o efeito podem implementar-se várias ações. Entre elas, acompanhar o volume do fluxo de comunicação para identificar se há ou não aumento das conversas *online*. Por exemplo, neste estudo de caso o acompanhamento identificou que, das 22 semanas analisadas, 3 registaram um aumento considerável das conversas *online*.

Pode também analisar-se a satisfação e/ou insatisfação dos clientes não só para quantificar a sua extensão, mas também para evidenciar o valor dos dados sociais. A identificação dos problemas dos clientes, não apenas numa escala qualitativa, mas também quantitativa, oferece às marcas informação das perceções e opiniões dos clientes. Essa informação pode ser útil para os departamentos de comunicação, imagem e relações públicas. A identificação das mensagens mais partilhadas e o seu conteúdo, para além de identificar palavras-chave também dá uma ideia do que os utilizadores andam a ver e em que RSO, o que transmite à marca como e onde deve participar. Por último, a identificação de utilizadores com poder, importantes, influentes, isto é, líderes do ponto de vista da SNA, ajuda a orientar o tipo de conteúdo e a que segmentos deve a marca chegar.

[70] Acesso em junho de 2019, disponível no portal do Observador: https://observador.pt/2019/01/14/novo-recorde-vw-vendeu-1083-milhoes-de-veiculos/.

Acesso em janeiro de 2017, disponível no portal do Observador: https://observador.pt/2016/12/18/escandalo-das-emissoes-qual-escandalo-europeus-nao-parecem-preocupados/.

Acesso em janeiro de 2017, disponível no portal da *Automotive News Europe*: https://europe.autonews.com/article/20161215/ANE/161219920/vw-s-cheat-diesels-hold-their-value-in-the-used-car-market.

6.5 Resumo

Os estudos de caso apresentados neste capítulo, para além de responderem a casos concretos de apoio à decisão, tiveram como objetivo definir o modelo apresentado neste trabalho de investigação. Neste capítulo não foi possível, por limitação de espaço, apresentar toda a informação quantitativa e qualitativa produzida. Também não foi possível representar todas as redes criadas e analisadas até os dados estarem totalmente refinados. Por esse motivo, apresentaram-se apenas os resultados mais relevantes e algumas redes criadas durante todo o processo.

O estudo de caso "Leitões" foi o primeiro passo, realizado para caracterizar e identificar as áreas de conhecimento e os recursos necessários para analisar os dados das RSO para apoio à decisão. Através dele abordou-se a complexidade subjacente à análise do discurso *web*, focaram-se os aspetos importantes e a considerar na sua análise, bem como a sua utilização no apoio à decisão no contexto das RSO. Exploraram-se também as ferramentas de extração e manipulação de dados das RSO. Este estudo de caso mostrou que as técnicas utilizadas e os resultados obtidos na análise do discurso *web* podiam ser utilizadas nas fases do processo de decisão.

No segundo estudo de caso, "TAP", abordou-se a complexidade subjacente à análise de grandes volumes de dados, recolha e processamento de dados temporais (*snapshots*). Abordaram-se também técnicas de visualização gráfica da SNA para *data-mining* e *text-mining*. Através do estudo de caso "TAP" identificou-se a necessidade de incorporar no modelo um algoritmo para PLN. O algoritmo, para além de executar as rotinas de limpeza dos dados não estruturados, transforma o conteúdo semântico dos *posts* em tantos pedaços quantos os conceitos existentes. O objetivo deste estudo de caso foi, não só executar uma análise estrutural da rede para identificar informação relevante, mas também explorar os conteúdos semânticos publicados para observar a satisfação e/ou insatisfação do cliente durante uma greve ocorrida em maio de 2015. Este estudo de caso mostrou que a estruturação das entidades do discurso e a análise semântica do conteúdo dos *posts*, fornecem informação útil para acompanhamento e apoio de decisões.

O terceiro estudo de caso, "Morangos", teve como objetivo resolver um problema de tomada de decisão, utilizando o modelo proposto neste trabalho de investigação, para extrair, processar, estruturar e analisar dados das RSO. Com a aplicabilidade, em dados reais, identificou-se a necessidade de incorporar no modelo não só redes semânticas de

palavras-chave, mas também redes semânticas para resumir o conteúdo das trocas discursivas. Este estudo de caso mostrou que os dados semânticos extraídos das trocas discursivas, podem ser convertidos em informação quantitativa para apoio à decisão.

O quarto e último estudo de caso, Volkswagen (VW), permitiu validar o modelo proposto. A recolha e análise dos dados da VW teve como objetivo perceber o impacto do escândalo de emissões de poluentes que foi divulgado em 18 de setembro 2015. Neste estudo de caso abordaram-se e exploraram-se as métricas e as técnicas de visualização gráfica da SNA no sentido de desenvolver indicadores para monitorizar as "conversas" *online*. Abordou-se também a complexidade da análise de grandes volumes de dados, recolha e processamento de dados temporais. Este estudo de caso mostrou que a monitorização das RSO pode melhorar e apoiar a tomada de decisão, no sentido de limitar eventuais danos de imagem.

7 Conclusões, limitações e trabalhos futuros

7.1 Conclusões

Este trabalho de investigação estruturou-se em duas partes. Na primeira, abordaram-se os temas teóricos chave da investigação, RSO, SNA e apoio à decisão e na segunda parte apresentou-se um modelo desenvolvido para análise de RSO e quatro estudos de caso. Os estudos de caso permitiram identificar e caracterizar os domínios teóricos e técnicos necessários para desenvolver o modelo, bem como aplicá-lo em diferentes contextos para estruturar, de forma simples, os dados das RSO, para extrair informação do discurso *web* para o apoio à decisão.

Este trabalho enquadrou-se na área da Ciência Aplicada à Decisão (CAD), mais especificamente na análise de RSO para apoio à decisão. Utilizaram-se os métodos formais da SNA, para representar e analisar as RSO. Interligaram-se também as áreas da computação, do PLN, linguística, etc., abordando-se a complexidade subjacente à estrutura dos dados do discurso *web*, para representar formalmente as ligações entre os três níveis de análise (utilizador, mensagem e conceito). Focaram-se ainda os aspetos importantes da teoria dos grafos e das técnicas da SNA a ter em consideração na sua análise, bem como na sua utilização para apoio à decisão no contexto das RSO.

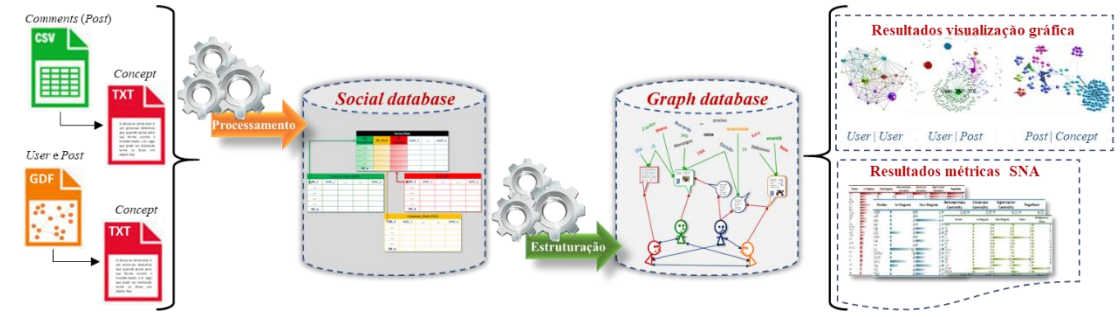
Os motivos que justificaram esta investigação em RSO para apoio à decisão baseiam-se nos factos de que a investigação existente é sobretudo teórica e a literatura encontra-se fragmentada pelas ciências sociais e da computação. Existe também a falta de evidências de como modelar e operacionalizar as interações entre utilizadores e os seus conteúdos semânticos. Além do mais, a diversidade dos conteúdos das RSO contém dados e informação relevantes.

A investigação atual depende de metodologias e tecnologias complexas que não permitem interligar e aceder em tempo útil aos dados sociais, para os medir ou monitorizar de forma eficaz. Persiste ainda o desconhecimento comum na interpretação e compreensão dos efeitos do volume dos dados, das suas características e diversidade de fontes. Por fim, as teorias e modelos tradicionais de apoio à decisão assentes em processos racionais e sequenciais onde se parte do pressuposto de que existe, *a priori*, um problema de decisão definido, para o qual são necessárias alternativas.

Adotou-se a metodologia *action design research*, devido ao facto da investigação envolver a definição de um modelo para implementar e responder a questões na área das RSO e apoio à decisão, e a sua criação implicar decidir sobre o tipo de dados necessários, a identificação de participantes e de fontes de dados, a definição de métodos de recolha e análise de dados, a contextualização da investigação e a definição de um período temporal. Em traços gerais, a metodologia permitiu que tanto o conhecimento, como a compreensão e a solução do problema, se obtivessem através da construção e da aplicação do modelo proposto (artefacto), criado através de um determinado número de fases, para um contexto específico. O artefacto desenvolveu-se utilizando a *action design research* de forma interativa e recursiva, ou seja, sempre que necessário, voltava-se atrás para aperfeiçoar ou corrigir algum problema identificado.

O modelo estruturou-se sequencialmente para otimizar as etapas do processo de análise, extração e processamento de dados das RSO e para a obtenção de *outputs* úteis para apoio à decisão num contexto organizacional específico. Como mostra a Figura 62, resumidamente: (i) os *inputs* de dados de quem comunica com quem (ficheiros GDF, CSV, TXT); (ii) são processados e estruturados numa configuração de rede (*graph database*); (iii) as ligações sociais estabelecidas entre utilizadores, bem como os conteúdos que publicaram *online*, são inferidos e representados (qualitativamente e quantitativamente) para apoio à decisão.

Figura 62 - Componentes do processo de análise de RSO



(i) *Input*: fontes de dados (ii) Processamento e estruturação de dados (iii) *Outputs*: análise de dados

O modelo incorporou técnicas consideradas complexas das ciências da computação, PLN e ciências dos dados. Todavia, caracteriza-se por ser de simples e fácil implementação, pois recorreu-se a *software* de utilização comum normalmente em ambiente empresarial. O modelo suporta-se em conhecimentos teóricos e práticos, ferramentas e metodologias de domínios multidisciplinares encadeados de forma recursiva, para responder à complexidade

e incerteza associadas ao processo de decisão em ambientes *online*. O objetivo da sua utilização é melhorar o entendimento e otimização dos dados de um problema de decisão, de um contexto específico, através da sua recolha, processamento e análise. Esta análise, possibilita o apoio à decisão, interligando os conteúdos das RSO com os resultados, quer quantitativos, quer qualitativos da SNA.

O modelo proposto desenvolveu-se e testou-se com recurso a quatro estudos de caso. O processo iniciou-se com a recolha de registos de interação social das RSO, seguindo-se o seu armazenamento, processamento, resultados e informação para apoio à decisão. Os dados, dos estudos de caso, extraíram-se diretamente de dois grupos e de duas *FanPages* do *Facebook* e o seu processamento executou-se não só com recurso a técnicas e métodos ligados às ciências sociais, mas também das ciências da computação. O modelo também integrou a limpeza, estruturação e padronização dos dados semânticos das mensagens, classificados na literatura como dados não estruturados. O modelo responde assim à questão da falta de recursos semânticos padronizados, de utilização generalizada para processamento dos conteúdos não estruturados de um contexto específico. A estruturação dos dados, interpretação e processamento semântico executaram-se com o auxílio de um algoritmo, desenvolvido para o efeito, para processamento dos mesmos e de uma *cleaning database* que auxiliou na sua limpeza.

Mais especificamente, o modelo proposto apreende a estrutura social do contexto em análise, oferece às organizações a oportunidade de integrar facilmente os dados sociais nas várias fases do processo de decisão, tem aplicabilidade em contextos organizacionais de pequena ou de grande dimensão, permite identificar quem interage com quem e ainda a visualização dos dados numa única rede ou da desagregação desta em sub-redes.

Através dos estudos de caso demonstrou-se não só o funcionamento do modelo, mas também se exemplificou a sua aplicabilidade com dados de contextos organizacionais reais. Os dados e os resultados finais traduziram a informação para apoio à tomada de decisão de forma compacta e visualmente atraente, através de métodos formais da SNA, matrizes e grafos, que representaram as características das interações sociais de forma sistematizada. As representações gráficas da rede, que englobaram os utilizadores, mensagens e conceitos, por um lado resumiram o discurso produzido nas RSO, enquanto por outro, representaram redes de palavras-chave. A sua visualização permitiu uma interpretação mais intuitiva dos dados dos três níveis de análise e ajudou a acelerar não apenas o seu processamento, mas também a análise semântica.

Através do estudo de caso "Leitões" percebeu-se que o discurso das RSO podia ser utilizado para apoio à decisão e que as trocas discursivas podiam ser medidas através das métricas da SNA. Concluiu-se que era possível ligar as etapas do processo de análise de RSO e da SNA com as quatro fases do processo de decisão e que os conteúdos produzidos pelos utilizadores ofereciam aos decisores informação útil e relevante.

O estudo de caso "TAP" teve dois objetivos. O primeiro, aplicar a SNA a grandes volumes de dados para identificar palavras-chave que resumissem o conteúdo do discurso e o segundo, verificar se os conteúdos semânticos, recolhidos em intervalos de tempo regulares, poderiam ser utilizados para apoio à decisão em contexto de crise. Concluiu-se que a utilização e o cálculo de métricas da SNA e a análise de múltiplas redes oferecia uma visão mais estruturada dos envolvidos no discurso (TAP e seus clientes) e dos conceitos mais utilizados. Concluiu-se, ainda, que a análise e a visualização dos conceitos podiam utilizar-se como *input* no acompanhamento e apoio à decisão, por exemplo, para decidir quais os canais a incluir ou a reforçar nas estratégias de comunicação, no sentido de diminuir o tempo de resposta aos clientes. Os resultados da análise semântica ofereceram informação útil das palavras-chave utilizadas que podiam ser usadas para analisar várias alternativas de um problema e para apoiar a tomada de decisões. Concluiu-se também que a análise dos conteúdos dos *posts* permitiria a sua classificação por categorias, o que permitiria uma melhor gestão dos recursos organizacionais para direcionar as solicitações.

No estudo de caso "Morangos" exploraram-se as redes de duas perspetivas diferentes: a análise das interações sociais entre utilizadores e a análise semântica do discurso produzido pelos mesmos. Concluiu-se que era possível extrair informação para apoio à decisão e que os conteúdos textuais continham informação semântica útil para melhorar o desempenho do apoio à decisão. Através deste estudo de caso, concluiu-se também que os dados semânticos (informação qualitativa) extraídos do discurso *web*, podem ser convertidos em informação quantitativa para apoio à decisão.

Por último, no estudo de caso "*Volkswagen*", o objetivo foi perceber as repercussões na reputação da marca VW, devido ao escândalo das emissões de gases. A análise dos conteúdos semânticos dos *posts* e a identificação de palavras-chave permitiram obter informação relevante dos tópicos em discussão e verificar que quem entrava na *FanPage* da VW ficava inevitavelmente exposto à história do escândalo. A SNA evidenciou conceitos mais negativos, mas também fez sobressair informação positiva. Uma interpretação possível foi que, independentemente de os utilizadores estarem preocupados

com as emissões de gases poluentes, a reputação da marca não foi afetada apesar de ter ficado exposta.

Em conclusão, os resultados obtidos reforçam que a extração de informação e conhecimento útil, a partir das interações entre utilizadores de RSO, é importante para o apoio à decisão. A análise dos dados sociais, em tempo útil, tem contrapartidas para as empresas pois a melhoria nas decisões podem refletir uma melhoria nos produtos e serviços, eventuais reduções de custos e aumentos de receita.

7.2 Limitações e trabalhos futuros

As características multidisciplinares e interdisciplinares desta área tornam-na extensa, logo as limitações identificadas encontram-se associadas a esta investigação e foi a partir delas que foi possível identificar oportunidades que podem vir a ser consideradas em trabalhos futuros.

Relativamente às limitações encontradas na execução deste trabalho de investigação foram essencialmente duas: as diretamente ligadas à linguagem e as associadas a questões práticas e técnicas.

Nas primeiras, destacam-se as especificamente relacionadas com a linguística e os conceitos utilizados, que podem ser diferentes de contexto para contexto. Para além disso, idiomas diferentes têm estruturas linguísticas diferentes. Por exemplo, no inglês, o adjetivo vem normalmente antes do substantivo (*delicious strawberries*) enquanto que no português o adjetivo vem habitualmente após o substantivo (morangos [*strawberries*] deliciosos [*delicious*]). A localização geográfica desempenha um papel importante pois cada país tem uma cultura própria, logo a utilização da linguagem nas RSO pode diferir.

As segundas dizem respeito ao processo de transformação semântica. As técnicas de PLN utilizadas para estruturar os dados semânticos incluíram apenas sequências *unigram*: palavras individuais separadas pelo algoritmo. Outra limitação identificada, envolve questões de escalabilidade do modelo e a capacidade de o sistema, em *Excel*, suportar acréscimos sucessivos de dados e consequentes tarefas de armazenamento e processamento. Embora essas tarefas não sejam complexas, a sua eficácia pode ficar mais reduzida devido

às limitações do *Excel*[71] (número total de linhas e colunas da folha de cálculo e número total de caracteres por célula).

Em investigações futuras, pondera-se incluir uma dimensão geográfica, ou seja, a localização da RSO e dos seus utilizadores. Nesta dimensão, seria interessante ver estudos comparativos de outros países e linguagens específicas do contexto, para destacar diferenças e semelhanças entre os procedimentos de extração semântica. A dimensão geográfica poderia ainda ajudar a avaliar e identificar diferenças nos processos e comportamentos de apoio à decisão em contextos geográficos distintos.

Associadas às questões técnicas, identificam-se duas linhas de orientação para futuras investigações. Na primeira, destacam-se as técnicas de PLN utilizadas pelo algoritmo para estruturar os dados que apenas incluem sequências *unigrams*: palavras individuais separadas pelo algoritmo. Em trabalhos futuros seria interessante melhorar o algoritmo de processamento de dados semânticos para que sejam utilizadas sequências de *n-gram* superior a 1. Na segunda, seria interessante explorar a implementação do modelo com bases de dados num servidor. Por exemplo, utilizar o *Microsoft SQL Server* para que todo o processamento (que consome recursos computacionais) fique do lado do servidor e com isso se possa aumentar a eficácia das tarefas de armazenamento e processamento de dados. Assim, no *Excel* a informação para análise e apoio à decisão ficaria acessível através de uma ligação à base de dados *SQL*.

Para além das oportunidades de investigação referidas, outras áreas interessantes, para utilização do modelo desenvolvido, incluem os sistemas de recomendação, os sistemas de reputação e a *sentiment analysis*. Estas áreas, que envolvem a investigação em métodos e algoritmos de identificação de palavras-chave e de deteção de tópicos, encontram-se em expansão, como observado no subcapítulo 3.4. Todavia, o mesmo não pode ser dito relativamente à sua interligação com as fases de recolha, armazenamento e processamento semântico dos dados, pois poucos estudos padronizam os seus conteúdos.

Nos sistemas de recomendação, seria interessante utilizar a informação semântica, após a aplicação do modelo, como *input* de qualquer algoritmo de recomendação, com o objetivo de identificar as preferências de clientes (utilizadores) e personalizar serviços com base nessa informação. Estes sistemas envolvem tarefas complexas, sobretudo devido à

[71] Especificações e limites do *Excel* para *Office 365*, *Excel* 2019, *Excel* 2016, *Excel* 2013, *Excel* 2010, *Excel* 2007. Acesso em junho de 2019, disponível no portal da *Microsoft*: https://support.office.com/pt-pt/article/especifica%C3%A7%C3%B5es-e-limites-do-excel-1672b34d-7043-467e-8e27-269d656771c3.

necessidade de considerar informação explícita e implícita, informação incompleta ou ausência da mesma, aspetos contextuais, entre outras contempladas no modelo.

Ainda nos sistemas de reputação, seria interessante utilizar o modelo para explorar, classificar e quantificar níveis de confiança, pois nestes sistemas, os padrões de influência são utilizados como referência para certificar (ou não) o nível de confiança dos utilizadores. Nesta área, a análise da disseminação de informação nas RSO seria útil para apreender como a mesma se propaga e chega a um maior número de clientes. Por exemplo, no *marketing online*, a ideia subjacente é de que utilizadores muito influentes podem ativar uma reação em cadeia, impulsionada pela consecutiva partilha de conteúdos e, assim, atingir grande parte da RSO.

Na *sentiment analysis*, o modelo poderia ser utilizado como metodologia de extração de informação, que capta a semântica contida nos conteúdos, utilizando conceitos predefinidos. O objetivo seria classificar de forma explícita conteúdos subjetivos, utilizando conjuntos definidos de conceitos. Para além disso, permitiria avaliar a polaridade (positiva, negativa ou neutra) das opiniões expressas nos conteúdos classificados como subjetivos.

Ainda no âmbito da *sentiment analysis* e utilizando como exemplo o estudo de caso "TAP", concluiu-se que, através da análise do conteúdo dos *posts*, os mesmos poderiam ser classificados por categorias. Essas categorias podem ser por tipo de mensagem, isto é, se é uma queixa, um esclarecimento sobre atraso no voo, uma informação sobre o cancelamento de um voo, entre outras. A categorização das mensagens permitiria uma melhor gestão dos recursos organizacionais para direcionar as solicitações. Para além disso, permitiria criar grupos de atendimento que respondessem assertivamente a casos de queixas, solicitações de apoio (por exemplo, suporte na alteração de uma reserva) ou de informação (sobre hotel, recolha de bagagem, procedimentos de reembolso, etc.).

Por último, esta investigação focou-se em quatro exemplos práticos e, apesar de terem sido utilizados dados reais, seria interessante a aplicação do modelo em mais contextos empresariais, assim como noutros tipos de RSO.

8 Referências

Abraham, A., Hassanien, A. E., Snásel, V. (2010). *Computational Social Network Analysis: Trends, Tools and Research Advances*. London, UK: Springer.

Adam, F. (2008). "Using Network Analysis for Understanding How Decisions are Made", in Adam, F., Humphreys, P. (eds.), *Decision Making and Decision Support Technologies*. New York: Idea Publishing Group, 950-957.

Aggarwal, C. C. (2011). *Social Network Data Analytics*. New York: Springer.

Agogo, D., Hess, T. J. (2018). "Scale Development Using Twitter Data: Applying Contemporary Natural Language Processing Methods in IS Research", in Deokar, A. V., Gupta, A., Iyer, L. S., Jones, M. C. (eds.), *Analytics and Data Science: Advances in Research and Pedagogy*. Switzerland: Springer, 163-178.

Al-Taie, M. Z., Kadry, S. (2017). *Python for Graph and Network Analysis*. Switzerland: Springer.

Aladwani, A. M. (2014). "The 6As model of social content management" *International Journal of Information Management*. 34 (2), 133-138.

Alexander, M., Kusleika, D. (2016). *Excel 2016 Power Programming with VBA*. Canada: Wiley.

Alhajj, R., Rokne, J. (2018). *Encyclopedia of Social Network Analysis and Mining*. New York: Springer.

Alkhyeli, M., Mansour, A. (2015). "Using Social Media for Supporting Decision-Making in Managing Public Relations: The Case of Abu Dhabi Police", in Mesquita, A., Peres, P. (eds.), *ECSM 2015 2nd European Conference on Social Media*. Porto: Academic Conferences and Publishing International, 479-487.

Antunes, F., Costa, J. P. (2011) "Decision Support Social Network" Conference in *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference*. Chaves, Portugal 15-18 de Junho de 2011.

Antunes, F., Costa, J. P. (2012a) "Disentangling Online Social Networking and Decision Support Systems Research Using Social Network Analysis", Liverpool, UK 12-13 de abril de 2012.

Antunes, F., Costa, J. P. (2012b). "Integrating Decision Support and Social Networks" *Advances in Human-Computer Interaction*. 2012, 1687-5893.

Antunes, F., Costa, J. P. (2012c) "The Interconnection of DSS and Online Social Networking" Conference in *Sistemas e Tecnologias de Informação: 7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. Madrid, Spain 20-23 de Junho de 2012.

Antunes, F., Freire, M., Costa, J. P. (2014). "Semantic Web Tools and Decision-making", in Zaraté, P., Kersten, G. E., Hernández, J. E. (eds.), *Group Decision and Negotiation: A*

Process-Oriented View, Lecture Notes in Business Information Processing (LNBIP). Switzerland: Springer, 270-277.

Antunes, F., Freire, M., Costa, J. P. (2016). "Semantic web and decision support systems" *Journal of Decision Systems*. 25 (1), 79-93.

Antunes, F., Freire, M., Costa, J. P. (2018). "From motivation and self-structure to a decision support framework for online social networks", in Ahuja, V., Rathore, S. (eds.), *Multidisciplinary Perspectives on Human Capital and Information Technology Professionals*. IGI Global.

Appel, A. P., Santana, V. F. d., Moyano, L. G., Ito, M., Pinhanez, C. S. (2018). A Social Network Analysis Framework for Modeling Health Insurance Claims Data, *Computer Science. Social and Information Networks*. arXiv Cornell University.Acesso em 24 de julho de 2020, disponível em Portal da Cornell University: https://arxiv.org/abs/1802.07116

Arif, T. (2015). "The Mathematics of Social Network Analysis: Metrics for Academic Social Networks" *International Journal of Computer Applications Technology and Research*. 4 (12), 889-893.

Atkinson, K., Bench-Capon, T., McBurney, P. (2006). "PARMENIDES: Facilitating Deliberation in Democracies" *Artificial Intelligence and Law*. 14 (4), 261-275.

Aufaure, M. A., Le Grand, B., Soto, M., Bennacer, N. (2006). "Metadata and Ontology Based Semantic Web Mining", in Taniar, D., R., J. W. (eds.), *Web Semantics and Ontology*. London, UK: Idea Group Publishing, 259-295.

Bahga, A., Madisetti, V. (2019). *Big Data Science & Analytics: A Hands-on Approach*. Arshdeep Bahga, Vijay Madisetti.

Balk, H. (2008). IMPACT Half Year Report: National Library of the Netherlands.

Banati, H., Bhattacharyya, S., Mani, A., Koppen, M. (2017). *Hybrid Intelligence for Social Networks*. Springer.

Banerjee, S., Jenamani, M., Pratihar, D. K. (2017) "Properties of a Projected Network of a Bipartite Network" *in 2017 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India 6-8 de abril de 2017.

Banica, L., Brinzea, V. M., Radulescu, M. (2015). "Analyzing Social Networks from the Perspective of Marketing Decisions" *Scientific Bulletin - Economic Sciences*. 14, 37-50.

Barão, A. (2014). Network Splitter 3D, Acesso em 24 de julho de 2020, disponível em Portal: http://baraoalexandre.blogspot.com/2014/10/gephi-plugin-network-splitter-3d.html

Barends, E., Rousseau, D. M. (2018). *Evidence-Based Management: How to Use Evidence to Make Better Organizational Decisions*. USA: Kogan Page.

Barlow, M. (2013). *Real-Time Big Data Analytics: Emerging Architecture*. USA: O'Reilly Media.

Bassan, A. S., Sarkar, D. (2014). *Mastering SQL Server 2014 Data Mining*. UK: Packt Publishing.

Bastian, M., Heymann, S., Jacomy, M. (2009) "Gephi: an open source software for exploring and manipulating networks" *in Third International ICWSM Conference*, San Jose, California USA 17-20 de maio de 2009.

Bechmann, A., Lomborg, S. (2015). *The Ubiquitous Internet. User and Industry Perspectives*. New York: Taylor & Francis.

Beck-Fernandez, H., Nettleton, D. F., Recalde, L., Saez-Trumper, D., Barahona-Peñaranda, A. (2017). "A system for extracting and comparing memes in online forums" *Expert Systems With Applications*. 82 (2017), 231-251.

Beer, D. (2009). "Power through the algorithm? Participatory web cultures and the technological unconscious" *New Media & Society*. 11 (6), 985-1002.

Benjamin, A., Chartrand, G., Zhang, P. (2017). *The Fascinating World of Graph Theory*. USA: Princeton University Press.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web, *Scientific American*. Acesso em 24 de julho de 2020, disponível em Portal da Scientific American: http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

Bhanap, S., Kawthekar, S. (2015). "Data Mining for Business Intelligence in Social Network: A survey" *International Advanced Research Journal in Science, Engineering and Technology* 2(12), 129-131.

Bharti, S. K., Pradhan, R., Babu, K. S., Jena, S. K. (2017). "Sarcasm Analysis on Twitter Data Using Machine Learning Approaches", in Missaoui, R., Abdessalem, T., Latapy, M. (eds.), *Trends in Social Network Analysis: Information Propagation, User Behavior Modeling, Forecasting, and Vulnerability Assessment* Switzerland: Springer, 51-76.

Biswas, S., Bordoloi, M., Shreya, J. (2018). "A graph based keyword extraction model using collective node weight" *Expert Systems With Applications*. 97 (2018), 51-59.

Bodomo, A. B. (2010). *Computer-Mediated Communication for Linguistics and Literacy: Technology and Natural Language Education*. Hershey, New York: Information Science Reference.

Bonacich, P. (1987). "Power and Centrality: a Family of Measures" *The American Journal of Sociology*. 92 (5), 1170-1182.

Bonacich, P. (2007). "Some unique properties of eigenvector centrality" *Social Networks*. 29 (2007), 555–564.

Bonchi, F., Castilho, C., Gionis, A., Jaimes, A. (2011). "Social Network Analysis and Mining for Business Applications" *ACM Transactions on Intelligent Systems and Technology (TIST)*.

Borgatti, S. P. (2009). "2-Mode Concepts in Social Network Analysis", in Meyers, R. A. (eds.), *Encyclopedia of Complexity and System Science*. California: Springer, 8279-8291.

Borgatti, S. P., Brass, D. J., Halgin, D. S. (2014). "Social network research: confusions, criticisms, and controversies", in Brass, D. J., Labianca, G., Mehra, A., Halgin, D. S., Borgatti, S. P. (eds.), *Contemporary Perspectives on Organizational Social Networks*. Bradford, UK: Emerald Publishing, 1-29.

Borgatti, S. P., Halgin, D. S. (2011). "Analyzing Affiliation Networks", in Scott, J., Carrington, P. J. (eds.), *The SAGE Handbook of Social Network Analysis*. UK, USA: Sage, 417-433.

Bouet, M., Gançarski, P., Aufaure, M. A., Boussaid, O. (2009). "Pattern Mining and Clustering on Image Databases", in Erickson, J. (eds.), *Database Technologies: Concepts, Methodologies, Tools, and Applications*. USA: IGI Global, 60-85.

Boyd, D. M., Ellison, N. B. (2007). "Social network sites: Definition, history, and scholarship" *Journal of Computer-Mediated Communication*. 13 (1), 210-230.

Brandes, U., Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Berlin Heidelberg: Springer.

Brins, S., Page, L. (1998). "The anatomy of a large-scale hypertextual web search engine" *Computer Networks and ISDN Systems*. 30 (1998), 107-117.

Bucciarelli, E., Chen, S., Corchado, J. M. (2018). *Decision Economics: In the Tradition of Herbert A. Simon's Heritage: Distributed Computing and Artificial Intelligence, 14th International Conference*. Switzerland: Springer.

Burguillo, J. C. (2018). *Self-organizing Coalitions for Managing Complexity: Agent-based Simulation of Evolutionary Game Theory Models using Dynamic Social Networks for Interdisciplinary Applications*. Switzerland: Springer.

Campbell, W. M., Dagli, C. K., Weinstein, C. J. (2013). "Social Network Analysis with Content and Graphs" *MIT Lincoln Laboratory Journal*. 20 (1), 62-81.

Carley, K. M. (2015). "Dynamic Network Analysis", in Breiger, R., Carley, K. M. (eds.), *in the Summary of the NRC workshop on Social Network Modeling and Analysis*. Pittsburgh, USA: National Research Council.

Caroleo, B., Tosatto, A., Osella, M. (2015). "Making Sense of Governmental Activities Over Social Media: A Data-Driven Approach", in Delibasié, B., Hernández, J. E., Papathanasiou, J., Dargam, F., Zaraté, P., Ribeiro, R., Liu, S., Linden, I. (eds.), *Decision Support Systems V – Big Data Analytics for Decision Making*. Belgrade, Serbia: Springer, 34-45.

Carrington, P. J., Scott, J., Wasserman, S. (2005). *Models and Methods in Social Network Analysis*. Cambridge, New York: Cambridge University Press.

Castrucci, M., Delli Priscoli, F., Pietrabissa, A., Suraci, V. (2011). "A Cognitive Future Internet Architecture", in Domingue, J., Galis, A., Gavras, A., Zahariadis, T., Lambert, D., Cleary, F., Daras, P., Krco, S., Müller, H., Li, M. S., Schaffers, H., Lotz, V., Alvarez, F.,

Stiller, B., Karnouskos, S., Avessta, S., Nilsson, M. (eds.), *The Future Internet 2011: Achievements and Technological Promises*. New York: Springer, 91-102.

Celko, J. (1999). *Joe Celko's Data and Databases: Concepts in Practice*. San Francisco: Morgan Kaufmann Publishers.

Chakraborty, S., Tripathy, B. K. (2016). "Alpha-anonymization techniques for privacy preservation in social networks" *Social Network Analysis and Mining*. 6 (29), 1-11.

Chatterjee, A., Trumbo, B. E. (2018). "Univariate Descriptive Statistics", in Alhajj, R., Rokne, J. (eds.), *Encyclopedia of Social Network Analysis and Mining*. New York: Springer, 3252-3272.

Chen, H., Chiang, R. H. L., Storey, V. C. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact" *MIS Quarterly*. 36 (4), 1165-1188.

Cherven, K. (2013). *Network Graph Analysis and Visualization with Gephi*. UK: Packt Publishing.

Cherven, K. (2015). *Mastering Gephi Network Visualization*. UK: Packt Publishing.

Chester, S., Kapron, B. M., Srivastava, G., Srinivasan, V., Thomo, A. (2018). "Anonymization and De-anonymization of Social Network Data", in Alhajj, R., Rokne, J. (eds.), *Encyclopedia of Social Network Analysis and Mining*. New York: Springer.

Croft, B. W., Metzler, D., Strohman, T. (2015). *Search Engines: Information Retrieval in Practice*. Pearson Education.

D'Andrea, A., Ferri, F., Grifoni, P. (2010). "An Overview of Methods for Virtual Social Networks Analysis", in Abraham, A., Hassanien, A. E., Snásel, V. (eds.), *Computational Social Network Analysis - Trends, Tools and Research Advances*. London, UK: Springer, 3-25.

Dale, R. (2010). "Classical Approaches to Natural Language Processing", in Indurkhya, N., Damerau, F. J. (eds.), *Handbook of Natural Language Processing* 2 ed. USA: Chapman & Hall/CRC.

Das, S. (2016) "Topics and relationship discovery from unstructured equivocal sources for situation assessment" Conference in *19th International Conference on Information Fusion (FUSION)*. Heidelberg, Germany 5-8 de julho de 2016.

Das, T. K., Teng, B.-S. (1999). "Cognitive Biases and Strategic Decision Processes: An Integrative Perspective" *Journal of Management Studies*. 36 (6), 757-778.

Davenport, T. H. (2014). *Big data at work: dispelling the myths, uncovering the opportunities*. USA: Harvard Business School Press.

Davis, A. (2019). *Data Wrangling with JavaScript*. New York: Manning Publications Company.

de Nooy, W., Mrvar, A., Batagelj, V. (2018). *Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software*. UK: Cambridge University Press.

Dey, N., Bhatt, C., Ashour, A. S. (2019). *Big Data for Remote Sensing: Visualization, Analysis and Interpretation*. Switzerland: Springer.

Diestel, R. (2017). *Graph Theory*. Berlin, Germany: Springer.

Duari, S., Bhatnagar, V. (2020). "Complex Network based Supervised Keyword Extractor" *Expert Systems With Applications*. 140 (2020).

Duijn, M. A. J., Vermunt, J. K. (2006). "What Is Special About Social Network Analysis?" *Methodology*. 2 (1), 2-6.

Dunaway, M. M. (2018). "An Examination of ERP Learning Outcomes: A Text Mining Approach", in Deokar, A. V., Gupta, A., Iyer, L. S., Jones, M. C. (eds.), *Analytics and Data Science: Advances in Research and Pedagogy*. Switzerland: Springer, 265-280.

Dziczkowski, G., Bougueroua, L., Wegrzyn-Wolska, K. (2010). "IA-Regional-Radio – Social Network for Radio Recommendation", in Abraham, A., Hassanien, A.-E., Snásel, V. (eds.), *Computational Social Network Analysis - Trends, Tools and Research Advances*. London, UK: Springer, 413-436.

Erétéo, G., Limpens, F., Gandon, F., Corby, O., Buffa, M., Leitzelman, M., Sander, P. (2011). "Semantic Social Network Analysis: A Concrete Case", in Daniel, B. K. (eds.), *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena*. USA: IGI Global, 122-138.

Everett, M. G., Borgatti, S. P. (2005). "Extending Centrality", in Carrington, P. J., Scott, J., Wasserman, S. (eds.), *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, 57-75.

Fairclough, N. (2003). *Analysing Discourse*. London, UK: Routledge.

Faust, K. (2005). "Using Correspondence Analysis for Joint Displays of Affiliation Networks", in Carrington, P. J., Scott, J., Wasserman, S. (eds.), *Models and Methods in Social Network Analysis*. New York Cambridge University Press, 117-147.

Feldman, R., Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. UK: Cambridge University Press.

Felt, M. (2016). "Social media and the social sciences: How researchers employ Big Data analytics" *Big Data & Society*. 3 (1), 2053951716645828.

Fernando, G., MdJohar, M., MdJohar, M. (2015). "Framework for Social Network Data Mining" *International Journal of Computer Applications Technology and Research*. 116 (18), 7-10.

Filip, F. G., Zamfirescu, C. B., Ciurea, C. (2017). *Computer-Supported Collaborative Decision-Making*. Switzerland: Springer.

Freeman, L. C. (1979). "Centrality in Social Networks Conceptual Clarification" *Social Networks*. 2 (3), 215-239.

Freeman, L. C. (2004). *The Development of Social Network Analysis: A study in the Sociology of Science*. Vancouver, Canada: BookSurge, LLC.

Freeman, L. C. (2005). "Graphic Techniques for Exploring Social Network Data", in Carrington, P. J., Scott, J., Wasserman, S. (eds.), *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, 248-269.

Freire, M., Antunes, F., Costa, J. P. (2015a). "Discurso Web, Modelos Teóricos e Sistemas de Argumentação: Implicações para a Tomada de Decisão, *Relatório de Investigação do INESC, n.º 2/2015*. Coimbra, Portugal.

Freire, M., Antunes, F., Costa, J. P. (2015b) "Exploring social network analysis techniques on decision support" *in 2nd European Conference on Social Media (ECSM 2015)*, Porto

Freire, M., Antunes, F., Costa, J. P. (2015c) "Social network analysis to support decision-making" Conference in *WAM 2015 - Workshop on Assessment Methodologies - energy, mobility and other real world applications* Coimbra, Portugal 19 de Junho de 2015.

Freire, M., Antunes, F., Costa, J. P. (2017). "A semantics extraction framework for decision support in context-specific social web networks", in Linden, I., Liu, S., Christian, C. (eds.), *Decision Support Systems VII. Data, Information and Knowledge Visualization in Decision Support*. Switzerland: Springer, 133-147.

Freire, M., Antunes, F., Costa, J. P. (2021). "Getting decision support from context-specific online social networks: a case study" *Submetido para revisão.*

Friedkin, N. E. (1991). "Theoretical Foundations for Centrality Measures" *American Journal of Sociology*. 96 (6), 1478-1504.

Fruchterman, T. M. J., Reingold, E. M. (1991). "Graph drawing by force-directed placement" *Software: Practice and experience*. 21 (11), 1129-1164.

Fu, X., Luo, J. D., Boos, M. (2017). *Social Network Analysis: Interdisciplinary Approaches and Case Studies*. USA: Taylor & Francis Group.

Gardner, R. (2004). "Conversation Analysis", in Davies, A., Elder, C. (eds.), *The Handbook of Applied Linguistics*. Oxford, UK: Blackwell, 262-284.

Gartner, R. (2016). *Metadata: Shaping Knowledge from Antiquity to the Semantic Web*. Switzerland: Springer.

Gee, J. P. (2001). *An Introduction to Discourse Analysis - Theory and Method*. New York: Taylor & Francis.

Georgalou, M. (2010). "Pathfinding Discourses of Self in Social Network Sites", in Taiwo, R. (eds.), *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*. New York: IGI Global, 39-62.

Ghafoor, F., Niazi, M. A. (2016). "Using social network analysis of human aspects for online social network software: a design methodology" *Complex Adaptive Systems Modeling*. 4 (14).

Ghim, G.-H., Kim, K., Ko, Y., Bae, S., Choi, W. (2018). "NetMiner", in Alhajj, R., Rokne, J. (eds.), *Encyclopedia of Social Network Analysis and Mining*. New York: Springer, 1450-1474.

Gjoka, M., Kurant, M., Butts, C. T., Markopoulou, A. (2011). "A Walk in Facebook: Uniform Sampling of Users in Online Social Networks" *IEEE INFOCOM (2010)*. San Diego, CA: IEEE Journal on Selected Areas in Communications (JSAC), 1-9.

Gliwa, B., Zygmunt, A. (2015). "Finding Influential Bloggers" *International Journal of Machine Learning and Computing*. 5 (2), 127-131.

Golbeck, J. (2015). *Introduction to Social Media Investigation: A Hands-on Approach*. USA: Syngress.

Gordon, T. F. (2011) "Combining Rules and Ontologies with Carneades" *in 5th International RuleML2011@BRF Challenge*, Barcelona, Spain 19-21 de julho de 2011.

Gou, L., You, F., Guo, J., Wu, L., Zhang, H. (2011) "SFViz: Interest-based Friends Exploration and Recommendation in Social Networks" *in JCDL '10 Proceedings of the 10th annual joint conference on Digital libraries*, Hong Kong, China

Grohe, M. (2017). *Descriptive Complexity, Canonisation, and Definable Graph Structure Theory*. UK: Cambridge University Press.

Guerrero, H. (2019). *Excel Data Analysis: Modeling and Simulation*. Switzerland: Springer.

Gyarmati, L., Trinh, T. A. (2009). "Characterizing User Groups in Online Social Networks", in Oliver, M., Sallent, S. (eds.), *The Internet of the Future*. Berlin, Heidelberg: Springer, 59-68.

Haj-Bolouri, A., Purao, S., Matti, R., Lennarth, B. (2018) "Action Design Research in Practice: Lessons and Concerns" Conference in *Twenty- Sixth European Conference on Information Systems (ECIS2018)*. Portsmouth,UK.

Hanneman, R. A., Riddle, M. (2005). Introduction to Social Network Methods. 2005. Acesso em 24 de julho de 2020, disponível em Portal da University of California, Riverside: http://faculty.ucr.edu/~hanneman/networks/nettext.pdf

Hansen, D. L., Shneiderman, B., Smith, B. (2011). *Analysing Social Media Networks with NodeXL: Insights From a Connected World*. Burlington, MA, USA: Morgan Kaufmann.

Harrison, G. (2015). *Next Generation Databases NoSQL, NewSQL, and Big Data*. New York, USA: Springer.

Hatamura, Y. (2006). *Decision-Making in Engineering Design: Theory and Practice*. London, UK: Springer.

He, J., Chu, W. W. (2010). "A Social Network - Based Recommender System (SNRS)", in Memon, N., Xu, J. J., Hicks, D. L., Chen, H. (eds.), *Data Mining for Social Network Data*. New York: Springer, 47-74.

Herring, S. C. (2004). "Computer-mediated discourse analysis: An approach to researching online behavior", in Barab, S. A., Kling, R., Gray, J. H. (eds.), *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press, 338-376.

Herring, S. C. (2010). "Web Content Analysis: Expanding the Paradigm", in Hunsinger, J., Klastrup, L., Allen, M. (eds.), *The International Handbook of Internet Research*. New York: Springer, 233-249.

Herring, S. C. (2011). Computer- Mediated Conversation: Introduction and Overview, *Language@Internet*, 8, article 2. Acesso em 24 de julho de 2020, disponível em Portal da language@internet: http://www.languageatinternet.org/articles/2011/Herring

Herring, S. C. (2013). "Discourse in Web 2.0: Familiar, Reconfigured, and Emergent", in Tannen, D., Trester, A. M. (eds.), *Discourse 2.0: language and new media*. Washington, DC: Georgetown University Press, 1-25.

Hevner, A. R. (2007). "A Three Cycle View of Design Science Research" *Scandinavian Journal of Information Systems*. 19 (2), 87-92.

Hevner, A. R., Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice*. New York: Springer.

Hevner, A. R., March, S., Park, J., Ram, S. (2004). "Design Science in Information Systems Research" *MIS Quarterly*. 28 (1), 75-106.

Heymann, S., Le-Grand, B. (2013) "Visual Analysis of Complex Networks for Business Intelligence with Gephi" Conference in *17th International Conference on Information Visualisation*. London, UK.

Himelboim, I., McCreery, S., Smith, M. (2013). "Birds of a Feather Tweet Together Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter" *Journal of Computer-Mediated Communication*. 18, 154–174.

Hollstein, B. (2014). "Mixed Methods Social Networks Research: An Introduction", in Domínguez, S., Hollstein, B. (eds.), *Mixed Methods Social Networks Research: Design and Applications*. New York: Cambridge University Press, 3-34.

Huber, G. P. (1981). "The Nature of Organizational Decision Making and the Design of Decision Support Systems" *Management Information Systems Quarterly*. 5 (2), 1-10.

Huisman, M., Van Duijn, M. A. J. (2011). "A Reader's Guide to SNA Software", in Scott, J., Carrington, P. J. (eds.), *The SAGE Handbook of Social Network Analysis*. UK, USA: Sage, 578-600.

Ignatov, D. I., Khachay, M. Y., Labunets, V. G., Loukachevitch, N., Nikolenko, S. I., Panchenko, A., Savchenko, A. V., Vorontsov, K. (2017). *Analysis of Images, Social Networks and Texts*. Switzerland: Springer.

Ikematsu, K., Murata, T. (2013). "A Fast Method for Detecting Communities from Tripartite Networks", in Jatowt, A., Lim, E.-P., Ding, Y., Miura, A., Tezuka, T., Dias, G., Tanaka, K., Flanagin, A., Dai, B. T. (eds.), *Social Informatics*. Switzerland: Springer, 192-205.

Ingersoll, G., Morton, T., Farris, D. (2013). *Taming Text: How to Find Organize and Manipulate It*. USA: Manning Publications.

Isson, J. P. (2018). *Unstructured Data Analytics: How to Improve Customer Acquisition, Customer Retention, and Fraud Detection and Prevention*. New Jersey, USA: Wiley.

Izquierdo, L. R., Hanneman, R. A. (2006). Introduction to the Formal Analysis of Social Networks Using Mathematical. 2006. Acesso em 24 de julho de 2020, disponível em Portal da WOLFRAM: https://library.wolfram.com/infocenter/MathSource/6638/

Jacomy, M., Venturini, T., Heymann, S., Bastian, M. (2014). "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software" *PLoS ONE*. 9 (6), 1-12.

Jamali, M., Abolhassani, H. (2006) "Different Aspects of Social Network Analysis" Conference in *Web Intelligence, 2006*. Hong Kong 18-22 de Dezembro de 2006

Jaworski, A., Coupland, N. (2006). *The Discourse Reader*. New York: Routledge.

Jay, J. (2017). *Excel VBA: Step-By-Step Guide to Learning Excel Programming Language for Beginners*. CreateSpace Independent Publishing Platform.

Jensen, C. S., Snodgrass, R. T., Soo, M. D. (1996). "Extending existing dependency theory to temporal databases" *IEEE Transactions on Knowledge and Data Engineering*. 8 (4), 563-582.

Johannesson, P., Perjons, E. (2014). *An Introduction to Design Science*. New York: Springer International Publishing.

Jokinen, K., Wilcock, G. (2017). *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*. Singapore: Springer.

Jorgensen, M., Phillips, L. (2002). *Discourse Analysis as Theory and Method*. London: SAGE.

Joseph, S. R., Letsholo, K., Hlomani, H. (2016). "Social Media Data Mining: An Analysis & Overview of Social Media Networks and Political Landscape" *International Journal of Database Theory and Application*. 9 (7), 291-296.

Jurafsky, D., Martin, J. H. (2009). *Speech and Language Processing: an introduction to natural language processing, computational linguistics and speech recognition*. Prentice Hall.

Kamki, J. (2016). *Digital Analytics: Data Driven Decision Making in Digital World*. Notion Press.

Kanagavel, R. (2019). *The Social Lives of Networked Students. Mediated Connections*. Switzerland: Palgrave Macmillan.

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. New Jersey, USA: Wiley.

Kaplan, A. M., Haenlein, M. (2010). "Users of the world unite! The challenges and opportunities of social media" *Business Horizons*. 53 (1), 59-68.

Kappas, A., Kramer, N. C. (2011). *Face-to-Face Communication over the Internet: Emotions in a Web of Culture, Language, and Technology*. Cambridge, UK: Cambridge University Press.

Karacapilidis, N. (2006). "An Overview of Future Challenges of Decision Support Technologies", in Gupta, J. N. D., Forgionne, G. A., Mora, M. (eds.), *Intelligent Decision-making Support Systems Foundations, Applications and Challenges*. London: Springer-Verlag, 386-398.

Kefi, H., Indra, S., Abdessalem, T. (2016) "Social media marketing analytics: a multicultural approach applied to the beauty & cosmetic sector" *in Pacific Asia Conference on Information Systems (PACIS)*, Chiayi, Taiwan

Kemper, C. (2015). *Beginning Neo4j Create relationships and grow your application with Neo4j*. New York, USA: Springer.

Khan, S., Ali, S. A., Hasan, N., Shakil, K. A., Alam, M. (2019). "Big Data Scientific Workflows in the Cloud: Challenges and Future Prospects", in Das, H., Barik, R. K., Dubey, H., Roy, D. S. (eds.), *Cloud Computing for Geospatial Big Data Analytics: Intelligent Edge, Fog and Mist Computing*. Switzerland: Springer.

Khatri, A., Schewenger, M. N., Finkelstein, N. (2014). "Big SQL 3.0 Functionality: A comprehensive approach to SQL-on-Hadoop" *International Journal of Computer and Information Technology (ijcit)*. 3 (6), 1121-1128.

Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (2017). *Advances in Knowledge Discovery and Data Mining*. Springer.

Kleminski, R., Kazienko, P. (2018). "Identifying Promising Research Topics in Computer Science", in Alhajj, R., Hoppe, H. U., Hecking, T., Brodka, P., Kazienko, P. (eds.), *Network Intelligence Meets User Centered Social Media Networks*. Switzerland: Springer, 231-241.

Kok, S., Rogers, R. (2016). "Rethinking migration in the digital age-transglocalization and the Somali diaspora" *Global Networks*, 1-24.

Kollwitz, C., Dinter, B., Krawatzeck, R. (2018). "Tools for Academic Business Intelligence and Analytics Teaching: Results of an Evaluation", in Deokar, A. V., Gupta, A., Iyer, L. S., Jones, M. C. (eds.), *Analytics and Data Science. Advances in Research and Pedagogy*. Switzerland: Springer, 227-250.

Koutra, D., Faloutsos, C. (2018). *Individual and Collective Graph Mining: Principles, Algorithms, and Applications*. Morgan & Claypool Publishers.

Kozinets, R. V. (2010). *Netnography: Doing Etnographic Research Online*. London, UK: SAGE.

Krempel, L. (2011). "Network Visualization", in Scott, J., Carrington, P. J. (eds.), *The SAGE Handbook of Social Network Analysis*. UK, USA: Sage, 559-577.

Kreuz, R. J., Caucci, G. M. (2007) "Lexical influences on the perception of sarcasm" *in Proceedings of the Workshop on Computational Approaches to Figurative Language*, New York

Krishna, R. P. M., Mohan, A., Srinivasa, K. G. (2018). *Practical Social Network Analysis with Python*. Springer.

Kroenke, D. M., Auer, D. J., Vandenberg, S. L., Yoder, R. C. (2017). *Database Concepts*. New York: Pearson Education.

Lai, L. S. L., To, W. M. (2015). "Content Analysis Of Social Media: A Grounded Theory Approach" *Journal of Electronic Commerce Research*. 16 (2), 138-152.

Laudon, K. C., Laudon, J. P. (2011). *Management Information Systems: Managing the Digital Firm*. New York: Prentice Hall.

Laumann, E., Marsden, P., Prensky, D. (1992). "The Boundary Specification Problem in Network Analysis", in Freeman, L. C., White, D. R., Romney, A. K. (eds.), *Research Methods in Social Network Analysis*. New Brunswick, USA: Transaction Publishers.

Leech, N., Collins, K. M., Onwuegbuzie, A. J. (2018). "Collecting Qualitative Data to Enhance Social Network Analysis and Data Mining", in Alhajj, R., Rokne, J. (eds.), *Encyclopedia of Social Network Analysis and Mining*. New York: Springer, 230-240.

Leydesdorff, L., Hellsten, L. (2006). "Measuring the Meaning of Words in Contexts: An automated analysis of controversies about 'Monarch butterflies ,' 'Frankenfoods,' and 'stem cells' " *Scientometrics*. 67 (2), 231-258.

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press.

Liu, Y., Safavi, T., Dighe, A., Koutra, D. (2018). "Graph Summarization Methods and Applications: A Survey" *ACM Computing Surveys*. 51 (3).

Livari, J., Venable, J. R. (2009) "Action research and design science research - Seemingly similar but decisively dissimilar" Conference in *European Conference on Information Systems (ECIS)*. Acesso em 24 de julho de 2020, disponível no Portal da Association for Information Systems: http://aisel.aisnet.org/ecis2009

Locke, T. (2004). *Critical Discourse Analysis*. London: Continuum International Publishing Group.

Lomax, H. T. (2004). "Discourse Analysis", in Davies, A., Elder, C. (eds.), *The Handbook of Applied Linguistics*. Oxford, UK: Blackwell, 133-164.

Lozares, C., López-Roldán, P., Bolibar, M., Muntanyola, D. (2015). "The structure of global centrality measures" *International Journal of Social Research Methodology*. 18 (2), 209-226.

Luftenegger, E., Comuzzi, M., Grefen, P. W. P. J. (2015). "Designing a tool for service-dominant strategies using action design research" *Service Business*. 11 (1), 161-189.

Lytras, M. D., Damiani, E., de Pablos, P. O. (2009). *Web 2.0 The Business Model*. New York: Springer.

Ma, H., Che, D. (2016). "An Integrative Social Network and Review Content Based Recommender System" *Journal of Industrial and Intelligent Information*. 4 (1), 69-75.

Madan, M., Chopra, M. (2015). "Using Mining Predict Relationships on the Social Media Network: Facebook (FB)" *International Journal of Advanced Research in Artificial Intelligence*. 4 (4), 60-63.

Markov, Z., Larose, D. T. (2007). *Data-mining the Web: Uncovering Patterns in Web Content, structure, and Usage*. Hoboken, NJ: John Wiley & Sons, Inc.

Marmo, R. (2011). "Web Mining and Social Network Analysis", in Zhang, H., Segall, R. S., Cao, M. (eds.), *Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications*. Hershey, PA: Information Science Reference, 202-211.

Marsden, P. (2002). "Egocentric and Sociocentric Measures of Network Centrality" *Social Networks*. 24 (2002), 407-422.

Marsden, P. V. (2005). "Recent Developments in Network Measurement", in Carrington, P. J., Scott, J., Wasserman, S. (eds.), *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

McColl, R., Ediger, D., Poovey, J., Campbell, D., Bader, D. A. (2014) "A Performance Evaluation of Open Source Graph Databases" *in First workshop on Parallel programming for analytics applications*, Orlando, Florida, USA

Meier, A., Kaufmann, M. (2019). *SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management*. Wiesbaden, Germany: Springer Vieweg.

Mesa, J. C., García, M. d. l. O. P., Jacinto, L. G. (2018). "Analysis of social innovation on social networking services" *European Journal of Social Work*. 21 (6), 902-915.

Mettler, T. (2015). "Contextu alizing a professiona l social network for health care: Experiences from an action design research study" *Info Systems Journal*, 1-24.

Mika, P. (2007). *Social Networks and the Semantic Web*. Barcelona, Spain: Springer.

Mintzberg, H., Raisinghani, D., Théorêt, A. (1976). "The Structure of "Unstructured" Decision Processes" *Administrative Science Quarterly*. 21 (2), 246-275.

Missaoui, R., Abdessalem, T., Latapy, M. (2017). *Trends in Social Network Analysis: Information Propagation, User Behavior Modeling, Forecasting, and Vulnerability Assessment*. Springer.

Moghaddam, A. S., Hosseinkhani, J., Chuprat, S., Birgani, A. M., Keikhaee, S. (2017). "Applying Social Network Analysis Techniques in Crawler Based Search Engine to Support

Web Terrorism Mining" *IJCSNS International Journal of Computer Science and Network Security*. 17 (8), 75-81.

Mojtaba, S., Bamakana, H., Nurgalievab, I., Qua, Q. (2019). "Opinion leader detection: A methodological review" *Expert Systems with Applications*. 115 (2019), 200-222.

Moreira, J. M., Carvalho, A. C., Horvath, T. (2019). *A General Introduction to Data Analytics*. USA: Wiley.

Morgado, F. (2016). *Programming Excel with VBA: A Practical Real-World Guide*. New York: Apress.

Morris, C. (1938). Fundamentos da Teoria dos Signos (Tradução de António Fidalgo). 1938. Acesso em 19 de março de 2014, disponível em Portal da Universidade da Beira Interior: http://bocc.ubi.pt/~fidalgo/semiotica/

Morville, P., Rosenfeld, L. (2006). *Information Architecture for the World Wide Web*. O'Reilly Media.

Moser, C., Groenewegen, P., Huysman, M. (2013). "Extending Social Network Analysis with Discourse Analysis: Combining Relational with Interpretive Data", in Özyer, T., Rokne, J., Wagner, G., Reuser, A. (eds.), *The Influence of Technology on Social Network Analysis and Mining*. NewYork: Springer, 547-561.

Moser, C., Hellsten, L., Groenewegen, P. (2011) "The Icing on the Cake: Combining Relational and Semantic Methods to Extract Meaning from Online Message Board Postings" *in Proceedings of the ACM WebSci'11*, Koblenz, Germany

Mullarkey, M. T., Hevner, A. R. (2018). "An elaborated action design research process model" *European Journal of Information Systems*, 6-20.

Nassar, M., Orabi, A. A.-R., Doha, M., Bouna, B. A. (2015) "An SQL-like Query Tool for Data Anonymization and Outsourcing" Conference in *International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*. London, UK 8-9 de junho de 2015.

Newman, M. E. J. (2005). "A Measure of Betweenness Centrality Based on Random Walks" *Social Networks*. 27 (1), 39–54.

Newman, M. E. J. (2006). "Modularity and community structure in networks" *PNAS*. 103 (23), 8577-8582.

Nisbet, R., Miner, G., Yale, K. (2018). *Handbook of Statistical Analysis and Data Mining Applications*. UK: Academic Press.

Niu, L., Lu, J., Zhang, G. (2009). *Cognition-Driven Decision Support for Business Intelligence: Models, Techniques, Systems and Applications*. Berlin Heidelberg: Springer.

Nutt, P. C., Wilson, D. C. (2010). *Handbook of Decision Making*. United Kingdom: John Wiley & Sons, Ltd.

Opsahl, T. (2013). "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients" *Social Networks*. 35 (2), 159-167.

Ortiz-Arroyo, D. (2010). "Discovering Sets of Key Players in Social Networks", in Abraham, A., Hassanien, A.-E., Snášel, V. (eds.), *Computational Social Network Analysis: Trends, Tools and Research Advances*. London, UK: Springer-Verlag.

Osei-Bryson, K.-M., Rayward-Smith, V. J. (2009). "Data mining and operational research: techniques and applications" *Journal of the Operational Research Society*. 60 (8), 1043–1044.

Oussalah, M., Bhat, F., Challis, K., Schnier, T. (2013). "A software architecture for Twitter collection, search and geolocation services" *Knowledge-Based Systems*. 2013 (37), 105-120.

Ozyer, T., Bakshi, S., Alhajj, R. (2019). *Social Networks and Surveillance for Society*. Switzerland: Springer.

Ozyer, T., Rokne, J., Wagner, G., Reuser, A. H. P. (2013). *The Influence of Technology on Social Network Analysis and Mining*. Vienna: Springer.

Page, L., Brin, S., Motwani, R., Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web: Technical Report n.º 422, Stanford InfoLab.

Panda, M., Dehuri, S., Wang, G. N. (2014). *Social Networking: Mining, Visualization, and Security*. New York: Springer.

Pang, B., Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Pennington, D. R. (2017). "Coding of Non-Text Data", in Sloan, L., Quan-Haase, A. (eds.), *The SAGE Handbook of Social Media Research Methods*. London: Sage, 232-250.

Perelman, C., Olbrechts-Tyteca, L. (2005). *Tratado da Argumentação: A Nova Retórica*. São Paulo: Martins Fontes.

Perkins, L., Redmond, E., Wilson, J. R. (2018). *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*. USA: Pragmatic Bookshelf.

Piedrahita, P., Borge-Holthoefer, J., Moreno, Y., González-Bailón, S. (2017). The contagion effects of repeated activation in social networks, *Social Networks*, 54, 326-335.

Pinheiro, C. A. R. (2011). *Social Network Analysis in Telecomunications*. John Wiley & Sons, Ltd.

Pippal, S., Batra, L., Krishna, A., Gupta, H., Arora, K. (2014). "Data mining in social networking sites: A social media mining approach to generate effective business strategies" *International Journal of Innovations & Advancement in Computer Science*. 3 (2), 22-27.

Póvoa, A. P. B., Corominas, A., Miranda, J. L. (2017). *Optimization and Decision Support Systems for Supply Chains*. Switzerland: Springer.

Power, D. J., Burstein, F., Sharda, R. (2011). "Reflections on the Past and Future of Decision Support Systems: Perspective of Eleven Pioneers", in Schuff, D., Paradice, D., Burstein, F.,

Power, D. J., Sharda, R. (eds.), *Decision Support: An Examination of the DSS Discipline*. New York: Springer, 25-48.

Power, D. J., Phillips-Wren, G. (2012). "Impact of Social Media and Web 2.0 on Decision-Making" *Journal of Decision Systems*. 20 (3), 249-261.

Pozzi, F. A., Fersini, E., Messina, E., Liu, B. (2017). *Sentiment Analysis in Social Networks*. USA: Morgan Kaufmann.

Price, D., Earn, D., Fisman, D., Dushoff, J. (2011) "Modelling and Analysis of Options for Controlling Persistent Infectious Diseases" *in BIRS - Banff International Research Station*, Alberta, Canada 28 fevereiro - 4 março de 2011.

Provost, F., Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. USA: O'Reilly Media.

Qasem, Z., Jansen, M., Hecking, T., Hoppe, H. U. (2017). "Influential Actors Detection Using Attractiveness Model in Social Media Networks", in Cherifi, H., Gaito, S., Quattrociocchi, W., Sala, A. (eds.), *Complex Networks & Their Applications V*. Switzerland: Springer, 123-134.

Raghunathan, B. (2013). *The Complete Book of Data Anonymization: From Planning to Implementation*. New York: CRC Press.

Rahman, S. (2017). *Basic Graph Theory*. Switzerland: Springer.

Rahwan, I., Zablith, F., Reed, C. (2007). "Laying the Foundations for a World Wide Argument Web" *Artificial Intelligence*. 171 (10-15), 897-921.

Rapp, Marino, A., Simeoni, R., Cena, F. (2017). "An ethnographic study of packaging-free purchasing: designing an interactive system to support sustainable social practices" *Behaviour & Information Technology*. 36 (11), 1193-1217.

Ravindran, S. K., Garg, V. (2015). *Mastering Social Media Mining with R*. Birmingham, UK: Packt Publishing.

Réda, S., Baba-Hamed, L., Djatou, A. (2017). "Twitter Social Networking for Recommendation", in Linden, I., Mareschal, B., Shaofeng, L., Papathanasiou, J., Colot, C. (eds.), *ICDSST 2017- 3rd International Conference on Decision Support System Technology: Data, Information and Knowledge Visualization in Decision Making*. Namur, Belgium: ICDSST, 161-168.

Rieder, B. (2012). "What is in PageRank? A Historical and Conceptual Investigation of a Recursive Status Index" *Computational Culture*. 2.

Rieder, B. (2013) "Studying Facebook via data extraction: the Netvizz application" Conference in *Proceedings of the 5th Annual ACM Web Science Conference*. Paris, France.

Riley, J. (2017). *Understanding Metadata: What is Metadata, and What is it For?* Baltimore: National Information Standards Organization (NISO).

Robinson, I., Webber, J., Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data*. Sebastopol, CA: O'Reilly.

Rodriguez, M. A., Steinbock, D. J., Watkins, J. H., Gershenson, C., Bollen, J., Grey, V., deGraf, B. (2007) "Smartocracy: Social Networks for Collective Decision Making" Conference in *40th Annual Hawaii International Conference on Systems Science - HICSS 2007*. Hawaii, USA 3-6 de janeiro de 2007

Roiger, R. J. (2017). *Data Mining: A Tutorial-Based Primer, Second Edition*. USA: CRC Press.

Rosenfeld, A., Kraus, S. (2018). *Predicting Human Decision-Making: From Prediction to Action*. Morgan & Claypool Publishers.

Roy, S. D., Mei, T., Zeng, W. (2014). "Bridging Human-Centered Social Media Content Across Web Domains", in Fu, Y. (eds.), *Human-Centered Social Media Analytics*. Boston, USA: Springer, 3-20.

Ruas, P. H. B., Machado, A. D., Silva, M. C., Meireles, M. R. G., Cardoso, A. M. P., Zárate, L. E., Nobre, C. N. (2019). "Identification and characterisation of Facebook user profiles considering interaction aspects" *Behaviour & Information Technology*. 38 (8), 858-872.

Rubin, A., Babbie, E. R. (2011). "Analyzing Existing Data: Quantitative and Qualitative Methods", in Rubin, A., Babbie, E. R. (eds.), *Research Methods for Social Work* 7th ed. Belmont, CA: Brooks/Cole Cengage Learning, 407-433.

Russell, M. A. (2014). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. USA: O'Reilly Media.

Russell, M. A., Klassen, M. (2019). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More*. USA: O'Reilly Media.

Sadovykh, V., Sundaram, D. (2016) "How Do Online Social Networks Support Decision Making? A Pluralistic Research Agenda" *in 49th Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA 5-8 de janeiro de 2016.

Sadovykh, V., Sundaram, D. (2017) "Decision Making Processes in Online Social Networks: A Comparative Analysis of Health and Financial Online Social Networks" *in 50th Hawaii International Conference on System Sciences*, Hawaii, USA 4-7 de janeiro 2017.

Sadovykh, V., Sundaram, D., Piramuthu, S. (2015a). "Do decision-making structure and sequence exist in health online social networks?" *Decision Support Systems*. 74 (2015), 102-120.

Sadovykh, V., Sundaram, D., Piramuthu, S. (2015b). "Do online social networks support decision-making?" *Decision Support Systems*. 70 (2015), 15-30.

Saint-Charles, J., Mongeau, P. (2018). "Social influence and discourse similarity networks in workgroups" *Social Networks*. 52, 228-237.

Saint-Dizier, P. (2012). "Processing Natural Language Arguments with the <TextCoop> Platform" *Argument & Computation*. 3 (1), 49-82.

Salamanos, N., Voudigari, E., Yannakoudakis, E. J. (2017). "Identifying Influential Spreaders by Graph Sampling", in Cherifi, H., Gaito, S., Quattrociocchi, W., Sala, A. (eds.), *Complex Networks & Their Applications V*. Switzerland: Springer, 111-122.

Samanthula, B. K., Jiang, W. (2014). "A Randomized Approach for Structural and Message Based Private Friend Recommendation in Online Social Networks", in Can, F., Ozyer, T., Polat, F. (eds.), *State of the Art Applications of Social Network Analysis*. Switzerland: Springer, 1-34.

Sapaty, P. S. (2019). *Holistic Analysis and Management of Distributed Social Systems*. Switzerland: Springer.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G. (2015). "Utilizing social media data for pharmacovigilance: A review" *Journal of Biomedical Informatics*. 54, 202-212.

Sathick, J., Venkat, J. (2015). "A Generic Framework for Extraction of Knowledge from Social Web Sources (Social Networking Websites) for an Online Recommendation System" *International Review of Research in Open and Distributed Learning*. 16 (2), 247-271.

Savic, M., Ivanovic, M., Jain, L. C. (2019). *Complex Networks in Software, Knowledge, and Social Systems*. Switzerland: Springer.

Scheuer, O., Loll, F., Pinkwart, N., McLaren, B. M. (2010). "Computer-Supported Argumentation: A Review of the State of the Art" *International Journal of Computer-Supported Collaborative Learning*. 5 (1), 43-102.

Schneider, J., Groza, T., Passant, A., Breslin, J. G. (2010). "A Review of Argumentation for the Social Semantic Web" *Semantic Web - Interoperability, Usability, Applicability*.

Scott, J. (2000). *Social Network Analysis: A Handbook*. London, UK: SAGE.

Sein, M. K., Henfridsson, O., Purao, S., Matti, R., Rikard, L. (2011). "Action Design Research" *MIS Quarterly*. 35 (1), 37-56.

Sheshasaayee, A., Jayanthi, R. (2015). "A Text Mining Approach to Extract Opinions from Unstructured Text" *Indian Journal of Science & Technology*. 8 (36), 1-4.

Shimazu, H., Shinichi, K. (2007). "KM2.0: Business Knowledge Sharing in the Web 2.0 Age" *NEC Technical Journal*. 2 (2), 50-54.

Shum, S. B. (2008) "Cohere: Towards Web 2.0 Argumentation" *in COMMA 2008 - Computational Models of Argument*, Toulouse, France 28-30 de maio de 2008.

Shum, S. B., Cannavacciuolo, L., De Liddo, A., Iandoli, L., Quinto, I. (2011). "Using Social Network Analysis to Support Collective Decision-Making Process" *International Journal of Decision Support System Technology*. 3 (2), 15-31.

Simon, H. A. (1977). *The New Science of Management Decision*. Upper Saddle River, New Jersey: Prentice Hall.

Sloan, L., Quan-Haase, A. (2017). *The SAGE Handbook of Social Media Research Methods*. USA: SAGE.

Smith, M. A., Hansen, D. L., Gleave, E. (2009a) "Analyzing Enterprise Social Media Networks" Conference in *Computational Science and Engineering*. Vancouver, BC 29-31 de agosto de 2009.

Smith, M. A., Shneiderman, B., Milic-Frayling, N., Rodrigues, E. M., Barash, V., Dunne, C., Capone, T., Perer, A., Gleave, E. (2009b) "Analyzing (Social Media) Networks with NodeXL" *in Proceedings of the Fourth International Conference on Communities and Technologies*, University Park, Pennsylvania, USA 25–27 de junho de 2009

Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C. (2018). "Social media analytics - Challenges in topic discovery, data collection, and data preparation" *International Journal of Information Management*. 39, 156-168.

Sueur, C., Deneubourg, J. L., Petit, O. (2012). "From Social Network (Centralized vs. Decentralized) to Collective Decision-Making (Unshared vs. Shared Consensus)" *PLos One*. 7 (2), 1-10.

Symeonidis, P., Ntempos, D., Manolopoulos, Y. (2014). *Recommender systems for location-based social networks*. New York: Springer.

Szabo, G., Polatkan, G., Boykin, O., Chalkiopoulos, A. (2018). *Social Media Data Mining and Analytics*. USA: Wiley.

Tannen, D. (2013). "The Medium Is the Metamessage: Conversational Style in New Media Interaction", in Tannen, D., Trester, A. M. (eds.), *Discourse 2.0: language and new media*. Washington, DC: Georgetown University Press, 99-117.

Tickoo, O., Iyer, R. (2017). *Making Sense of Sensors: End-to-End Algorithms and Infrastructure Design from Wearable-Devices to Data Centers*. New York: Apress.

Tiroshi, A., Kuflik, T., Berkovsky, S., Kaafar, M. A. (2017). Graph Based Recommendations: From Data Representation to Feature Extraction and Application, *Computer Science. Social and Information Networks*. arXiv Cornell University.https://arxiv.org/abs/1707.01250

Tollinen, A., Jarvinen, J., Karjaluoto, H. (2012). "Opportunities and Challenges of Social Media Monitoring in the Business to Business Sector" *The 4th International Business and Social Science Research Conference*. Dubai, UAE: 1-14.

Torgo, L. (2017). *Data Mining with R: Learning with Case Studies, Second Edition*. USA: CRC Press.

Torres-Moreno, J. M. (2014). *Automatic Text Summarization*. USA: Wiley.

Tripathy, B. K., Baktha, K. (2018). *Security, Privacy, and Anonymization in Social Networks: Emerging Research and Opportunities: Emerging Research and Opportunities*. USA: IGI Global.

Tripathy, B. K., Thakur, S., Chowdhury, R. (2017). "A Classification Model to Analyze the Spread and Emerging Trends of the Zika Virus in Twitter", in Behera, H. S., Mohapatra, D. P. (eds.), *Computational Intelligence in Data Mining*. Singapore: Springer, 643-650.

Troisi, O., Grimaldi, M., Loia, F., Maione, G. (2018). "Big data and sentiment analysis to highlight decision behaviours: a case study for student population" *Behaviour & Information Technology*. 37 (10-11), 1111-1128.

Turban, E., Aronson, J. E., Liang, T.-P. (2007). *Decision Support Systems and Intelligent Systems*. New Delhi: Prentice-Hall.

Turban, E., Outland, J., King, D., Lee, J. K., Liang, T.-P., Turban, D. C. (2018). *Electronic Commerce 2018 - A Managerial and Social Networks Perspective*. Switzerland: Springer.

Van Dijk, T. A. (1977). *Text and Context: Explorations in the. Semantics and Pragmatics of. Discourse*. London: Longman Linguistic.

Van Dijk, T. A. (1985). "Introduction: Levels and dimensions of discourse analysis", in Van Dijk, T. A. (eds.), *Handbook of Discourse Analysis*. London: Academic Press, 1-11.

Van Dijk, T. A. (2001). "Critical Discourse Analysis", in Schiffrin, D., Tannen, D., Hamilton, H. E. (eds.), *Handbook of Discourse Analysis*. Oxford, UK: Blackwell, 352-371.

Van Duijn, M. A. J., Vermunt, J. K. (2006). "What Is Special About Social Network Analysis?" *Methodology*. 2 (1), 2-6.

Vasconcelos, J. B., Barão, A. (2017). *Ciência dos dados nas organizações*. Lisboa: FCA.

Velde, B. v. d., Meijer, A., Homburg, V. (2015). "Police message diffusion on Twitter: analysing the reach of social media communications" *Behaviour & Information Technology*. 34 (1), 4-16.

Vercellis, C. (2011). *Business Intelligence: Data Mining and Optimization for Decision Making*. UK: Wiley.

Vicario, M., Zollo, F., Caldarelli, G., Scala, A., Quattrociocchi, W. (2017). "Mapping social dynamics on Facebook: The Brexit debate" *Social Networks*. 50, 6-16.

Villegas, E. B. (2016). "Facebook and its Disappearing Posts: Data Collection Approaches on Fan-Pages for Social Scientists" *The Journal of Social Media in Society*. 5 (1), 160-188.

Vosecky, J., Jiang, D., Leung, K. W.-T., Xing, K., NG, W. (2014). "Integrating Social and Auxiliary Semantics for Multifaceted Topic Modeling in Twitter" *ACM Transactions on Internet Technology (TOIT) - Special Issue on Foundations of Social Computing*. 14 (4), 1-24.

Wachsmuth, H. (2015). *Text Analysis Pipelines: Towards Ad-hoc Large-Scale Text Mining*. Switzerland: Springer.

Walha, A., Ghozzi, F., Gargouri, F. (2017) "ETL4Social-Data: Modeling Approach for Topic Hierarchy" *in 9th International Joint Conference on Knowledge Discovery,*

Knowledge Engineering and Knowledge Management (KEOD 2017), Madeira, Portugal 1-3 de novembro de 2017.

Walton, C. D. (2007). *Agency and the Semantic Web*. New York: Oxford University Press.

Wasserman, S., Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New-York, USA: Cambridge University Press.

Wodak, R., Meyer, M. (2009). "Critical Discourse Analysis: History, agenda, theory, and methodology", in Wodak, R., Meyer, M. (eds.), *Methods of Critical Discourse Analysis* 2 ed. London: Sage, 1-33.

Wooffitt, R. (2005). *Conversation Analysis and Discourse Analysis - A Comparative and Critical Introduction*. London: Sage.

Xhafa, F., Barolli, L., Barolli, A., Papajorgji, P. (2015). *Modeling and Processing for Next-Generation Big-Data Technologies*. Switzerland: Springer.

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods*. London: SAGE Publications.

Zadrozny, P., Kodali, R. (2013). *Big Data and Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources*. New York: Apress.

Zaidi, F., Muelder, C., Sallaberry, A. (2018). "Analysis and Visualization of Dynamic Networks", in Alhajj, R., Rokne, J. (eds.), *Encyclopedia of Social Network Analysis and Mining*. New York: Springer.

Zhai, C. X., Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. USA: Association for Computing Machinery and Morgan & Claypool Publishers.

Zheng, Q., Skillicorn, D. (2017). *Social Networks with Rich Edge Semantics*. Minneapolis, Minnesota, USA: CRC Press.

Zielinski, A., Middleton, S. E., Tokarchuk, L. N., Wang, X. (2013) "Social media text mining and network analysis for decision support in natural crisis management" *in International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Baden-Baden, Germany 12-15 de maio de 2013.