UNIVERSIDADE D
COIMBRA

João Manuel Nunes Marques

Análise do Impacto de Condições de Saúde nas Redes Sociais

Dissertação no âmbito do Mestrado Integrado em Engenharia Biomédica orientada pelo Professor Doutor Hugo Ricardo Gonçalo Oliveira e apresentada ao Departamento de Física da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Julho de 2019

Agradecimentos

Gostaria de começar por agradecer ao Professor Orientador deste trabalho, Doutor Hugo Gonçalo Oliveira, por toda a orientação e colaboração, desde o início ao fim do trabalho, as reuniões semanais e a disponibilidade permanente para me ajudar, bem como toda a paciência nos momentos em que me demonstrei mais atrapalhado. Foi sem dúvida um privilégio ser orientado pelo Professor, homem de muitas virtudes. Não encontro o mínimo defeito a apontar. Um muito sincero obrigado.

Queria também agradecer à Doutora Maria Margarida Martins Gonçalo e à Doutora Bárbara Roque Ferreira, médicas dermatologistas do Serviço de Dermatologia do Hospital da Universidade - Centro Hospitalar e Universitário de Coimbra, pela disponibilidade em nos receber e pela orientação dada sobre as doenças a incluir neste estudo.

Ao nível mais pessoal, quero agradecer a Deus, Cristo Jesus, meu sentido, minha esperança, minha vida, meu tudo, e aos santos homens e mulheres, cujo testemunho me inspira a querer dar mais. Quero agradecer à minha família, por todo o esforço e sacrifício. Aos meus amigos, por todo o ânimo que recebi. Aos Portugueses, a quem muito devo, pelo privilégio de poder estudar, porque as propinas não cobrem todos os custos necessários. Estou em dívida para com todos.

Resumo

As redes sociais, dentro das quais o Twitter, fazem hoje parte da vida de muitas pessoas. A possibilidade de partilhar opiniões e momentos revolucionou a imersão dos utilizadores: já não são meros espetadores de conteúdo, são também criadores. As redes sociais tornaram-se uma praça pública virtual, onde os utilizadores se exprimem. A disponibilidade pública de tantos dados incentiva investigadores a aprofundar uma variedade de ciências pela extração a informação que contêm. No caso da saúde, há várias modalidades a serem desenvolvidas, como a monitorização de epidemias e abrangência geoespacial, a análise de sentimentos/emoções, prevenção de comportamentos de risco, etc. Este projeto visa, em primeiro lugar, apurar se o autor de um tweet é portador da doença a que se refere, para depois se proceder à análise de emoções, necessária para avaliar o impacto psicológico e social de portar tal doença, e compreender o espectro emocional característico de cada patologia analisada. Neste trabalho serão abordadas algumas doenças em maior profundidade: as psicodermatoses ou dermatoses psicofisiológicas. São doenças de pele que beneficiam significativamente, para seu tratamento, do acompanhamento psicológico. A sintomatologia pode ser agravada pelo estado emocional, de stress ou ansiedade, podendo este ser também originado pelo estigma social causado pela aparência e consequente baixa autoestima. Serão utilizadas tecnologias de processamento de linguagem natural em texto e algoritmos de aprendizagem computacional, bem como o recurso a um modelo representativo de emoções. Os tweets recolhidos são constituídos de texto em língua portuguesa.

Palavras-Chave

Classificação de Texto, Análise de Emoções, Análise de Tópicos, Condições de Saúde, Redes Sociais.

Abstract

The social networks, such as Twitter, are today part of many people lives. The possibility to sharing opinions and moments has revolutionized the immersion of users: they are no longer mere spectators of content, they are also creators. Social networks have become a virtual public square where users express themselves. The public availability of so much data encourages researchers to delve into a variety of sciences by extracting the information they contain. In the case of health, there are several modalities being developed, such as epidemics monitoring and geospatial coverage, analysis of feelings / emotions, prevention of risk behaviors, etc. This project aims, first of all, to predict if the tweet author is a carrier of the disease to which he refers, then proceed to the analysis of emotions, necessary to evaluate the psychological and social impact of carrying such a disease, and understand the spectrum emotional character of each pathology analyzed. In this work, some diseases will be approached in greater depth: the psychodermatosis or psychophysiological dermatoses. These are skin diseases that benefit significantly from psychological treatment. The symptomatology can be aggravated by the emotional state, stress or anxiety, which can also be caused by the social stigma caused by the appearance and consequent low self-esteem. Natural language processing technologies and computational learning algorithms will be used, and a representative model of emotions. The tweets collected are written in Portuguese.

Keywords

Text Classification, Emotion Analysis, Topic Modeling, Health Condition, Social Networks.

Conteúdo

1 Introdução **1**

2 Fundamentação Teórica **5**
2.1 Representação da Emoção 5
2.2 Processamento de Linguagem Natural 7
2.3 Classificação de Texto 10
2.3.1 Classificação supervisionada 10
2.3.2 Modelos de Classificação 11
2.3.3 Avaliação do Desempenho da Classificação 13
2.3.4 Active Learning 14
2.4 Latent Dirichlet Allocation 14
2.5 Trabalhos Relacionados 14

3 Definição de Objetivos **17**
3.1 Obtenção do conjunto de dados 19
3.2 Pré-processamento do texto 20
3.3 Primeiras Análises aos Dados Recolhidos 23
3.3.1 Análise de Tópicos 23
3.3.2 Emoções 26
3.3.3 Distribuição de tweets no tempo 28
3.3.4 Comparação Portugal-Brasil 30
3.4 Redefinição de Objetivos 32

4 Classificação de tweets com base em Crowdsourcing **33**
4.1 Estrutura e resultados da anotação 34
4.2 Obtenção dos classificadores 38
4.3 Resultados da filtragem com classificador binário 41
4.3.1 Análise de Tópicos 42
4.3.2 Emoções 44
4.3.3 Distribuição dos tweets no tempo 45
4.4 Resultados da filtragem com três classificadores 50
4.4.1 Análise de Tópicos 50
4.4.2 Emoções 54
4.4.3 Distribuição dos tweets no tempo 55
4.5 Discussão 59

5 Classificação de tweets com base em Active Learning **61**
5.1 Implementando Active Learning 61
5.2 Resultados da anotação 62
5.3 Resultados da filtragem com classificador binário 66
5.3.1 Análise de Tópicos 66
5.3.2 Emoções 69

5.3.3 Distribuição dos tweets no tempo 70
5.4 Resultados da filtragem com três classificadores 71
5.4.1 Análise de Tópicos 71
5.4.2 Emoções 74

6 Discussão e Conclusão **75**
6.1 Sobre os Resultados 77
6.2 Trabalho Futuro 79

Lista de Abreviaturas

AL Active Learning

ASD Autor-Sentido-Doente

CS Crowdsourcing

IDF Inverse Document Frequency

ILA Individual-Literal-Autor: o autor é individual, o sentido é literal, o doente é o autor. Relativo à classificação de tweets.

LDA Latent Dirichlet Allocation

PLN Processamento de Linguagem Natural

SVC SVM para Classificação

SVM Suport Vector Machine

TF-IDF Term Frequency-Inverse Document Frequency

TF Term Frequency

Capítulo 1

Introdução

As redes sociais são cada vez mais um local privilegiado para partilha de opinião e outras informações, muitas vezes de cariz pessoal, pela população em geral. Entre as várias informações partilhadas encontram-se as condições de saúde, isto é, doentes procuram e partilham informação sobre as suas doenças, os seus sintomas, tratamentos, e ainda as suas preocupações, acabando por, implicitamente, transmitir o impacto emocional da doença nas suas vidas. No domínio da saúde, a rede social *Twitter*[1] apresentam-se como uma valiosa fonte de informação para analisar o impacto de várias doenças na população e obter algumas métricas relacionadas com a saúde pública. Há exemplos, tais como a investigação acerca da doença cardiovascular (Sinnenberg et al., 2016), um meio para apurar a má administração de antibióticos (Scanfeld et al., 2010) e identificar efeitos adversos na toma de medicamentos (O'Connor et al., 2014), bem como ferramentas para prever e rastrear surtos de gripe e outras epidemias em tempo-real (Aramaki et al., 2011, Santos and Matos, 2014). Este parece ser realmente o caminho futuro: desenvolver metodologias que permitam um autêntico cuidado de saúde, uma medicina de prevenção, onde se pretende manter e promover o estado saudável da população. Um passo para além do atual cuidado na doença, onde a pessoa procura assistência apenas quando se encontra doente. Episódios recentes, tais como surtos de legionella em Portugal[2][3] motivam também neste sentido da monitorização das redes sociais poder revelar dados determinantes para a avaliação da propagação de epidemias. Ainda mais, se os conteúdos partilhados nas redes sociais que mencionam uma doença ou um surto forem sujeitos a uma análise emocional, poderemos evidenciar se existe autenticidade no carácter urgente, através da presença da emoção negativa de alta intensidade.

Este trabalho pretende ir ao encontro desta premissa e tem como objetivo principal compreender de que forma as pessoas reagem no Twitter às suas condições de saúde, nomeadamente através do reconhecimento de emoções presentes nas publicações que mencionam as doenças em estudo. O Twitter fornece um meio para a recolha de publicações, através da pesquisa pela introdução de palavras-chave. É possível obter publicações que referem doenças específicas e outros aspetos relacionados à saúde, o que, só por si, pode dar uma ideia acerca das doenças mais mencionadas ou de palavras frequentemente associadas a estas doenças. No entanto, esta recolha não é perfeita, já que, muitas vezes, as condições são referidas num contexto promocional (p.ex., *"melhore a sua psoríase em*

[1] `https://twitter.com/`

[2] "Portugal teve 233 casos de legionela em 2017"
`https://observador.pt/2018/01/29/portugal-teve-233-casos-de-legionela-em-2017/`

[3] 2014: "OMS considera surto de Legionella em Portugal como "grande emergência de saúde pública"
`https://www.publico.pt/2014/11/11/sociedade/noticia/oms-considera-surto-de-legionaella-em-portugal-como-grande-emergencia-de-saude-publica-1675862`

apenas um mês") ou com um sentido metafórico (p.ex., *"estas notícias causam-me urticária!"*). Porém, recorrendo a técnicas de Processamento de Linguagem Natural (PLN) e Aprendizagem Computacional, será possível, a certo ponto, identificar as mensagens efetivamente publicadas por doentes. Com o foco nessas publicações, e mais uma vez recorrendo a técnicas de PLN, será possível apurar opiniões acerca da vivência com a doença ou da satisfação com o tratamento, na forma de espectro emocional. Especificamente, e ao contrário da maior parte dos trabalhos relacionados, este é focado em publicações escritas em português, ou seja, os resultados serão essencialmente válidos para países de expressão portuguesa, principalmente o Brasil, o maior e também o com mais atividade no Twitter, e ainda Portugal.

O foco do estudo proposto reside em quatro dermatoses[4]: psoríase, urticária, dermatite atópica, e vitiligo. Esta escolha deve-se essencialmente à relação conhecida destas doenças com o estado psicológico (Duncan and Koo, 2018), pois os seus sintomas tendem a agravar em situações de stress, ansiedade, ao ponto do acompanhamento psicológico se tratar de um aliado fundamental para suavizar os sintomas. Contudo, acreditamos que os métodos a usar possam ser aplicados a outras doenças, o que desejamos comprovar com a realização de um estudo semelhante com outras doenças bastante comuns, a asma e a diabetes. Olhando para as palavras associadas às doenças torna-se possível encontrar referências a medicamentos, o que permitirá tirar conclusões acerca da sua popularidade e comparar o espectro emocional de diferentes pares doença-medicamento. Esta é mais uma das possibilidades que decorrem do sucesso destes estudos.

O conjunto de dados é constituído por tweets[5] obtidos de forma automática, a partir da API[6] disponibilizada pelo próprio Twitter. O número de exemplares ronda um milhão. A partir das datas obtidas juntamente com os textos, é apresentado o número de publicações ao longo do tempo. Para perceber os assuntos mais populares presentes nos conjuntos de textos foi utilizada uma implementação de Latent Dirichlet Allocation (LDA), uma técnica de *topic modeling*. Recorrendo à plataforma online Figure-Eight[7], foi obtida a anotação de cerca de 500 tweets, realizada por pessoas, sobre diferentes aspetos, tais como a perceção do tipo de autor e a emoção transmitida. Com estes dados foram treinados algoritmos de classificação supervisionada, para catalogar o conjunto total de tweets de forma automática. De forma a aumentar o conjunto de treino, foram estimados os tweets mais significativos a serem anotados pelo autor deste documento, a partir da implementação de *active learning*. Uma vez previstas as características dos tweets, foram mapeados no referencial de Russell (Russell (1980)), onde a emoção é representada por duas coordenadas, valência e ativação. A análise foi feita considerando 3 níveis possíveis para ambas, resultando em 9 combinações possíveis, que não representam uma só emoção, mas um grupo de emoções.

Entre as conclusões retiradas do trabalho, confirmou-se que grande parte das publicações não é da autoria de doentes; verifica-se uma alta proporção de tweets que mencionam urticária no sentido figurado; a maior popularidade de vitiligo entre as dermatoses abordadas, devido à sua visibilidade, amplificada por figuras públicas que sofrem desta doença; independentemente da doença, a carga emocional da maior parte das publicações é negativa, próxima da raiva, frustração ou medo; a importância dos dias comemorativos para a sensibilização/informação nas redes sociais (p.ex., Dia Mundial da Psoríase); a distribuição temporal tende a revelar sazonalidade.

[4]Dermatose: nome genérico das doenças da pele.
[5]Nome frequentemente dado a publicações na rede social Twitter.
[6]`https://help.twitter.com/pt/rules-and-policies/twitter-api`
[7]`https://www.figure-eight.com/`

À presente introdução segue-se o capítulo de fundamentação teórica, **capítulo 2**, onde se abordam conceitos fundamentais para a compreensão do resto do documento: começa por abordar modelos de representação de emoções e ferramentas simples para aplicar em textos, passando para as tarefas comuns de PLN, a classificação de texto recorrendo aos métodos mais tradicionais. O mesmo capítulo termina com uma breve apresentação de vários trabalhos relacionados nesta área de estudo, nomeadamente aqueles que relacionam saúde e redes sociais.

No **capítulo 3** são relatadas as primeiras experiências que se seguem à obtenção do conjunto de dados de estudo: a disposição ao longo do tempo, a análise de vocabulário e a obtenção de distribuição emocional a partir de léxicos. Estas experiências mostraram ainda que a mera recolha de tweets que mencionam o nome da doença é um processo ruidoso, já que a doença pode ser referida no sentido figurado ou ser mencionada por alguém que não sofre dela.

Identificados problemas de ruído no conjunto de tweets, são traçados novos objetivos explicitados no **capítulo 4**, onde se projeta uma solução de classificação de texto baseada numa anotação inicial de textos com recurso a *crowdsourcing*, tanto para a filtragem de tweets como para a categorização emocional. Depois da aplicação dos classificadores a todo o conjunto de dados, é realizada uma análise mais fina ao vocabulário e à distribuição emocional ao longo do tempo, a partir de duas modalidades distintas de filtrar os tweets de interesse.

Na impossibilidade de se recolherem mais tweets anotados, é descrita no **capítulo 5** a implementação de uma solução de anotação de tweets pelo autor deste documento, selecionados com base em *active learning* de forma a anotar tweets que oferecem informação significativa ao conjunto de dados. Após ser enriquecido com um maior número de tweets, são repetidas as mesmas experiências anteriores e realizadas novas análises aos resultados.

Segue-se a discussão, onde se analisam os resultados anteriores, de forma específica por cada temática. Concluindo-se com uma breve apreciação, sugerem-se novos desafios e tarefas a ter em atenção em futuros estudos semelhantes.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são descritos os fundamentos teóricos deste trabalho, começando pelo modelo de representação de emoções adoptado, seguido de algumas tarefas de Processamento de Linguagem Natural, métodos supervisionados para a classificação de texto, *Active Learning* e análise de tópicos (LDA). Finalmente, segue ainda uma breve descrição de trabalhos relacionados.

2.1 Representação da Emoção

Existem vários modelos para a representação de emoções. Um dos modelos é conhecido por modelo categórico de Ekman-Friesen (Ekman and Friesen (1971)), que define um conjunto de seis emoções básicas: tristeza, raiva, surpresa, medo, nojo e felicidade.

Outro modelo proposto é denominado 'Modelo Circunflexo de Afeto', por Russell (1980). O autor propôs a representação de emoções numa base referencial, onde as duas dimensões fundamentais correspondem à 'valência', que se trata da variável utilizada em Análise de Sentimento, e à 'ativação'. A valência corresponde a uma grandeza qualitativa do sentimento, variando de negativa a positiva, i.e., dentro da gama do sentimento muito desagradável e um sentimento muito agradável. A ativação traduz a amplitude da expressão do sentimento. Por exemplo, um sentimento negativo (valência negativa) pode ser expresso com ativação alta e assim temos, por exemplo, a raiva; ou ser expresso com ativação baixa e assim temos, p.ex., a tristeza. Analogamente, para o sentimento positivo (valência positiva) temos a felicidade (ou Alegria) quando a ativação é alta, ou seja, uma reação entusiasmada; ou uma emoção serena ou relaxada, quando a expressão é contida (ativação baixa). As emoções referidas (Relaxado, Feliz, Triste e Zangado) são exemplos de cada quadrante do referencial bidimensional. O estado neutro corresponde à origem do referencial. Outra variável pode ser adicionada ao modelo: a Dominância, que trata de representar o quanto a situação na causa da emoção está sob o controlo da pessoa. Dada a extensão limitada dos blocos de texto analisados, os tweets, não será considerada a dominância na análise de emoções porque se trata de uma variável mais difícil de prever. O modelo originalmente proposto dá conta de inúmeras emoções. Outros autores, posteriormente, têm alterado o número e disposição de emoções no referencial. Por essa razão, neste trabalho será operada uma simplificação por quadrantes e fronteiras onde a emoção se inclui. Na literatura inglesa, as variáveis de valência, ativação e dominância são referidas por *valence*, *arousal* e *dominance*, respetivamente.

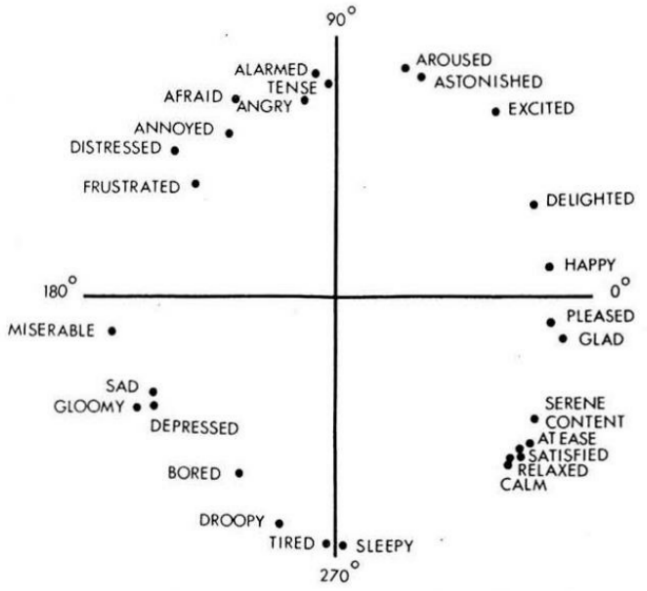

Figura 2.1: Modelo Circunflexo de Afeto, Russell (1980).

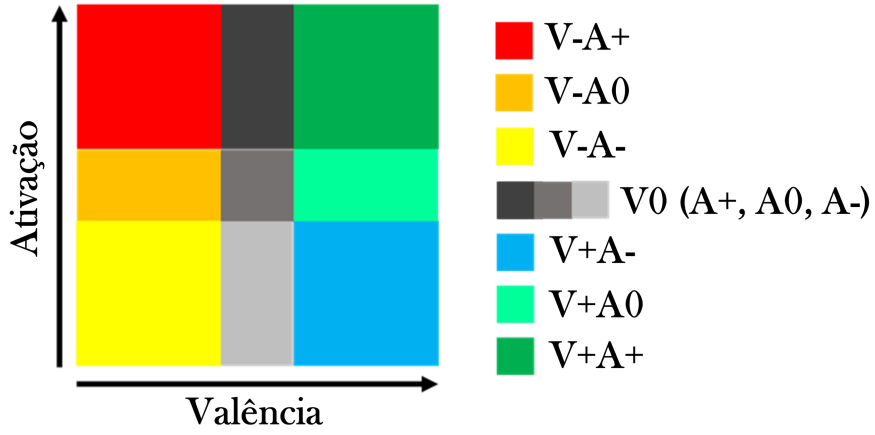

Figura 2.2: Modelo de representação de emoções adotado, baseado no Modelo circunflexo de Russel.

Uma forma simples de aplicar este tipo de modelo de representação de emoções é recorrendo a **Léxicos VAD**. Para obter a emoção representativa de um texto (um tweet, neste trabalho), é necessário analisar o seu conteúdo, essencialmente palavras e emojis/emoticons. Existem léxicos tanto de palavras como de emojis/emoticons, onde para cada elemento são apresentados os valores de valência e ativação. Contabilizando os elementos presentes num texto, são consideradas as coordenadas de Valência, Ativação e Dominância (VAD), e é feita uma ponderação para chegar à classificação final do tweet.

Neste trabalho será utilizado o léxico ANEW-PT para palavras, obtido por Soares et al. (2012). O léxico foi obtido da adaptação do original ANEW (Bradley and Lang, 1999) para o português europeu. Vários termos distintos em inglês com valores de valência, ativação e dominância diferentes, corresponderam ao mesmo termo em português. Dada a situação, foi realizada uma média de valores para os termos não únicos em português.

Para o caso de emojis/emoticons, o léxico utilizado será o LEED (Rodrigues et al., 2018a). Este léxico inclui os mesmos emojis de diferentes plataformas, com pequenas diferenças nos gráficos e nas cores, mas também nos valores das coordenadas emocionais. À semelhança do léxico para palavras, foi feita uma ponderação nestes casos. As variantes correspondentes ao Twitter (rede social em causa neste estudo) não foram incluídas, por isso os valores não puderam ser utilizados diretamente. Na tabela 2.1 apresenta-se um exemplo ilustrativo da utilização exclusiva de um léxico de emojis para classificar o seguinte tweet:

`Mano! Com apenas 1 estresse mnh psoríase atacou tudo!`💔😪👎

Emoji	Alias	Valência	Ativação
💔	Broken heart	0.11	0.84
😪	Sleepy face	0.21	0.60
👎	Thumbs down sign	0.17	0.61
	MÉDIA	0.16	0.68

Tabela 2.1: Exemplo ilustrativo da aplicação do léxico para classificar um texto que inclui os emojis listados, calculando a média. Os valores estão normalizados entre 0 e 1. O valor 0.5 separa a valência negativa (<0.5) da positiva (>0.5), bem como a ativação baixa (<0.5) da alta (>0.5).

Obtendo uma valência global negativa (neste caso, numericamente inferior a 0.5), e uma ativação alta, o tweet é localizado na região vermelha da figura 2.2, que inclui emoções como raiva, irritação.

2.2 Processamento de Linguagem Natural

A sigla 'NLP' é utilizada em língua inglesa como abreviação do termo 'Natural Language Processing'. Em língua portuguesa, 'Processamento de Linguagem Natural' é abreviado para 'PLN'. O objetivo do Processamento da Linguagem Natural passa por obter uma interação pessoa-máquina semelhante à interação pessoa-pessoa. Aliando a capacidade de compreender o discurso humano ao poder computacional, torna-se possível processar grandes quantidades de discurso humano, escrito ou falado, para uma grande variedade de fins. Designa-se por Linguagem Natural a linguagem própria dos humanos, nomeadamente a falada que pode ser expressa em forma escrita, palavra por palavra. O conceito é traçado em oposição às linguagens técnicas que os humanos têm de dominar para interagir com as máquinas. Sendo assim, o Processamento de Linguagem Natural tem parte na ciência informática com a finalidade de habilitar as máquinas à compreensão e produção de discurso, falado ou escrito, no jeito humano. O sucesso destas tarefas está dependente de vários fatores, como a definição lógica da estrutura do discurso, a sintaxe, ou o significado conceptual - a semântica, como o reconhecimento do contexto do discurso em ordem de desambiguar palavras homófonas.

Na tabela 2.2 podemos organizar o estudo linguístico em seis análises distintas, segundo Jurafsky and Martin (2009).

Quando partimos para o processamento de texto, já não é requerida uma análise fonética e fonológica. Mas interessa ainda fazer uma análise ao nível da morfologia se quisermos, por exemplo, consultar a palavra num dicionário. No dicionário não são apresentadas todas as variações de uma palavra. Frequentemente, são apresentados dois: o

Categoria	Objecto	Descrição
Fonética e Fonológica	Sons	Estuda a relação entre palavras e sons respectivos
Morfológica	Palavras	Estuda como as palavras são construídas (Morphe = forma)
Sintática	Frase	Estuda a relação entre palavras numa frase e a relação entre frases que compõem uma maior
Semântica	Conceitos	Estuda a relação entre as palavras e os seus significados
Pragmática	Contexto	Estuda como o uso das frases em contextos distintos adquire também significados distintos
Discursiva	Texto	Estuda a estrutura e o significado de um agrupamento de frases, a forma como influenciam o significado umas das outras

Tabela 2.2: Seis análises do estudo linguístico, segundo Jurafsky and Martin (2009).

substantivo simples ou o verbo no infinitivo, se aplicável, ou outra categoria de palavras como advérbios ou pronomes. Para consultar o dicionário é necessário reduzir a palavra a uma forma padrão, que varia com a função que está a desempenhar na frase, e que designamos por *lema*. São necessárias várias etapas de processamento para obter o *lema*, nomeadamente: Tokenização, Part-of-Speech Tagging e, finalmente, a Lematização.

Tokenização (*Tokenization*) - Segmentação de texto Esta tarefa corresponde à divisão do texto em *tokens*, i.e., isolar cada palavra e símbolo de pontuação, geralmente usando como delimitadores os espaços e os sinais de pontuação presentes. A palavra 'partiu-se' fica dividida em três tokens: 'partiu', '-' e 'se'.

Part-of-speech Tagging Para compreender o contexto em que um termo se insere, é necessário compreender a estrutura da frase. Esta operação tem como objetivo identificar qual a função desempenhada pela palavra (classe gramatical) ou símbolo no discurso: nome, pontuação, preposição, verbo infinitivo, verbo finito, advérbio, adjetivo, pronome, artigo, numeral. Assim como pode ser utilizado para identificar a classe gramatical das palavras na frase, também pode identificar as funções sintáticas: sujeito, predicado e complemento. Em termos práticos, a tarefa consiste em atribuir uma etiqueta (*tag*) a cada token, identificando a classe gramatical. Exemplo: 'Eu' → 'Eu#pronome-pessoal'.

Lematização (*Lemmatization*) Entende-se por *lema* a forma normalizada da palavra ou termo ao jeito de dicionário, i.e., a forma não flexionada, também designada de forma canónica. Os substantivos são geralmente convertidos para a forma masculina e singular, enquanto os verbos conjugados são convertidos para a forma verbal no infinitivo. Esta operação pode ser conseguida tendo em conta os pares $(token, tag)$ obtidos anteriormente, por exemplo: : "Expliquei" → tag: #verbo-finito → "Explicar".

Remoção de Stopwords Existem palavras no texto que não adicionam informação relevante em determinados processos de classificação ou análise, como a extração de palavras-chave. Remover estas palavras alivia o custo computacional no processamento do texto. Pronomes, verbos auxiliares, advérbios, artigos, preposições e pronomes constam geralmente nestas listas. Em alguns casos, a remoção de stopwords também pode ser nociva. Por exemplo, se é pretendido treinar um classificador que identifique se um autor se refere a si mesmo no texto ou não, os pronomes são essenciais para compreender a referência. As listas de stopwords são vulgarmente designadas por stoplists.

N-gramas Da mesma forma que pode ser interessante analisar a frequência de uma determinada palavra num texto, pode também ser relevante perceber o nível de ocorrência de certos agrupamentos de palavras. A utilização de uma dada palavra num texto, pode não revelar muito. No entanto, se as palavras que a rodeiam forem tidas em consideração, alguma ordem é conservada. Imagine-se um classificador de opinião. Extrair features semelhantes a "não é mau", pode ser de grande benefício. Agrupamentos de dois elementos designam-se de bigramas e agrupamentos de três elementos designam-se de trigramas.

TF-IDF

A forma mais simples de fornecer um texto a um sistema de classificação, é convertê-lo num saco de palavras (*bag-of-words*). Ou seja, trata-se de representar um texto na forma de vector-linha, onde cada coluna corresponde a uma palavras e onde está contido um valor. Esse valor pode ser binário (presente, não-presente), pode ser de frequência (número de vezes que a palavra ocorre no texto), pode ser de frequência relativa (TF, 'term frequency', número de vezes que a palavra ocorre a dividir pelo número de palavras, incluindo repetições, presentes), ou conter outro tipo de informação, totalmente personalizável.

TF-IDF é uma medida que relaciona a palavra de um documento, não apenas com as restantes palavras contidas nele, mas com as suas ocorrências em documentos diferentes. Por exemplo, numa tarefa de classificação de texto, se todos os documentos contêm o artigo 'a', este tende a ser uma *feature* irrelevante, podendo ser adicionada a uma lista de *stopwords*.

$$\text{tf-idf(t,d)} = \text{tf(t,d)} \times \text{idf(t)}$$

Existem variantes associadas ao cálculo do valor tf-idf. Na expressão acima, podemos ler "O valor tf-idf do termo t no documento d, é igual ao produto da frequência do termo t no documento d, com o inverso da frequência nos documentos do termo t".

O objetivo do primeiro factor (tf) é aumentar o valor do termo, quanto maior for a sua presença no documento d, por exemplo, um tweet. Já o objetivo do segundo factor (idf) é atenuar o valor do termo, quanto maior for a sua presença noutros documentos.

A versão mais simples do cálculo é multiplicar o número de vezes que o termo aparece no documento com $\log \frac{n}{\text{df}(t)}$, o logaritmo do quociente entre o número total de documentos (n) e o número total de documentos que contêm o termo.

Existem variantes: $tf(t, d)$ pode ser contabilizado como a frequência absoluta do termo t no documento d ou como a frequência relativa. O factor $idf(t)$ também admite vários tipos, dependendo da atenuação que se deseja.

Neste trabalho, quando for designada a contabilização dos termos por *Count* tratar-se-á da contagem absoluta de um termo no tweet. Quando for designada por TF consistirá na contagem relativa.

Algumas aplicações de PLN

No que toca ao estado atual de diferentes aplicações de PLN, podemos considerar a partir de um fórum de discussão online[1], que as tarefas de *Part-Of-Speech Tagging*, análise de dependências e sintaxe, conseguem ter um desempenho muito bom; o reconhecimento de entidades nomeadas (NER, *Named Entity Recognition*) atinge uma medida F1 de 0.8 a 0.9 para pequenas ontologias, mas quando aplicada a muitas classes, o desempenho mostra-se bastante limitado; no que toca a aplicações que geram respostas automáticas (*Question Answering*), é possível responder a questões básicas de uma base de conhecimento simples; no que toca a preenchimento de campos (*Slot filling*) como, por exemplo, um utilizador pedir a um assistente de voz para definir um alarme, o sistema consegue identificar com sucesso os campos específicos que deve programar: data, hora, repetição; da tarefa de tradução automática (*Machine translation*) conseguem-se obter resultados admissíveis; o reconhecimento de voz funciona bem se a voz for bem gravada sem ruído e sem sobreposição.

[1] `https://www.quora.com/What-is-the-state-of-art-in-Natural-Language-processing-as-of-2018`

2.3 Classificação de Texto

Nesta secção, apresenta-se algumas considerações sobre métodos supervisionados de classificação, seguidas de uma breve descrição de quatro algoritmos e medidas de avaliação de desempenho. Por fim, é apresentado a ideia base de active learning, com explicações mais detalhadas no decorrer do trabalho.

2.3.1 Classificação supervisionada

A classificação supervisionada trata de desenvolver um sistema que permita prever a classe de uma determinada amostra, partindo primeiro da dedução das relações entre as características de um determinado conjunto de amostras e as suas classes previamente anotadas. Reconhecendo os padrões entre features e classe anotada, o algoritmo adequa-se de forma a obter um conjunto de previsões o mais semelhante à realidade. Distingue-se da classificação não-supervisionada que, não contabilizando quaisquer classes de dados previamente anotados, separa as amostras na medida em que as suas features se distanciam, tentando reconhecer grupos de amostras naturalmente distintos entre si. Em termos práticos, para operar uma classificação supervisionada, é fornecido primeiramente ao sistema um conjunto de dados anotados. O conjunto contém as amostras e a cada amostra está correspondida uma classe anotada. Têm de existir no mínimo dois valores distintos para uma dada classe, já que o pretendido é obter um sistema que separe dados com classes diferentes. Este conjunto de dados anotados é denominado "Conjunto de Treino".

De seguida, o conjunto de treino é fornecido a um algoritmo que pode inspecionar as relações entre features e entre features e classes, e devolve um sistema com determinada eficácia em separar dados. Para testar a eficiência deste classificador é prudente aplicá-lo a um conjunto de dados novos com estrutura semelhante ao conjunto de treino, o "conjunto de teste", e comparar os resultados previstos com os resultados reais. Para este efeito, existe um conjunto numeroso de métricas, entre as quais, a precisão e *recall* (abrangência) e, a ponderação entre as duas, a medida F1.

Também pode ser aplicada uma validação cruzada. Existindo várias variantes, descreve-se uma delas.

O objetivo passa por apurar se o modelo utilizado incorre em *overfitting*, ou seja, se tende a apresentar bons resultados porque, perdendo generalidade, se adaptou demasiado aos dados de treino.

O conjunto de treino é assim dividido, p.ex., em 10 partes, e fazem-se 10 treinos, cada um excluindo uma das subpartes e sendo testado sobre ela. As métricas de desempenho são obtidas para cada uma das fases e é feita uma ponderação final sobre elas.

A explicação anterior trata das tarefas mais imprescindíveis no desenvolvimento de um classificador treinado segundo a abordagem supervisionada. Para obter um classificador eficaz é necessário primeiramente, obter um "conjunto de treino" consistente, com uma construção rigorosa, realizar uma análise mais detalhada aos dados e inserir tarefas intermédias, como o processamento de dados, extração de features, a seleção de features com base no contributo que cada uma pode adicionar ao sistema de classificação e a eliminação de features redundantes, bem como a utilização de técnicas como a Análise de Componentes Principais, que trata de transformar o espaço das features de forma a aumentar a obter o maior número de variância possível.

2.3.2 Modelos de Classificação

Os algoritmos de classificação usados neste trabalho, são provenientes de uma biblioteca para linguagem Python: **Scikit-Learn**[2]. As seguintes notas são baseadas na informação presente no site.

Naive Bayes, versão Multinomial

O classificador *Multinomial Naive Bayes* (`MultinomialNB`) é aplicável a problemas de classificação que usam features discretas. No caso de classificação de texto, por exemplo, as features podem corresponder à contagem de cada palavra num texto. Embora a distribuição multinomial exija, normalmente, números inteiros como features, na prática, números decimais como TF-IDF podem também funcionar.

O modelo aplica o teorema de Bayes admitindo de forma "ingénua" (*naive*) que todas as features são condicionalmente independentes entre si, e resulta da generalização do problema binomial. A cada classe y faz-se corresponder um vector $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$, onde n é o número de features (em classificação de texto será o tamanho do vocabulário) e θ_{yi} é a probabilidade $P(x_i \mid y)$ da feature i aparecer numa amostra que pertence à classe y.

Os valores que compõem o vector θ_y, na versão implementada, são obtidos segundo a fórmula de frequência relativa, modificada para prevenir probabilidades iguais a zero:

$$\hat{\theta}_{yi} = \frac{N_{yi}+1}{N_y+n},$$

onde o N_yi é o número de vezes que a feature i aparece numa amostra do conjunto de treino com classe y, e N_y é o somatório de todos os valores N_yi $(i = 1, 2, ..., n)$.

Support Vector Machines

Support vector machines (SVMs) são um conjunto de métodos supervisionados também aplicáveis em problemas de classificação. A ideia passa por criar um hiperplano (ou hiperplanos) para separar as instâncias de classes diferentes, aumentando o mais possível a margem de distância.

Este tipo de modelo lida bem com espaços de dimensionalidade muito alta. É ainda bastante versátil, admitindo diferentes *kernels* (linear, RBF, polinomial).

Neste trabalho é utilizada a implementação `LinearSVC`, de kernel linear. À sua semelhança, existe também um modelo destinado a tarefas de regressão, denominado `LinearSVR`.

A figura 2.3 consta da documentação da biblioteca Scikit-Learn[3] e ilustra o problema a duas dimensões.

[2] `https://scikit-learn.org/`

[3] `https://scikit-learn.org/stable/modules/svm.html`

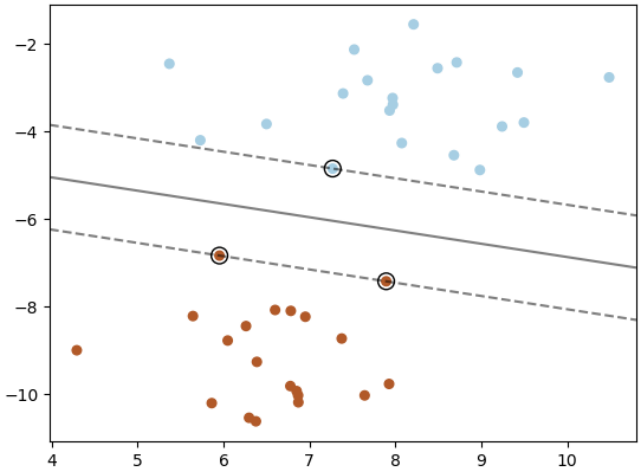

Figura 2.3: SVM: Figura ilustrativa a duas dimensões. O "hiper-plano" é aqui ilustrado por uma recta e as margens por tracejado.

Nearest Centroid

Este tipo de classificador é um método simples que representa cada classe com o seu centróide. Para cada classe, são reunidas as instâncias correspondentes incluídas no conjunto de treino. Depois, para cada feature, são considerados os valores presentes em cada uma das instâncias. É depois encontrado o ponto dentro dessa gama de valores, para o qual a soma das distâncias dos valores das instâncias a esse ponto é a menor, segundo determinada métrica (média ou mediana). Cada um destes pontos obtidos, consiste numa coordenada do centróide no hiperespaço de n-dimensões, onde n corresponde ao número de features.

Uma nova instância será classificada segundo o centróide que lhe é mais próximo. Como de trata de uma distância no espaço multidimensional, podem ser utilizadas diferentes métricas.

A medida utilizada neste trabalho foi a distância 'Euclidiana', definição-padrão na implementação usada, que calcula a distância entre dois pontos $A = (a_1, a_2, ..., a_n)$ e $B = (b_1, b_2, ..., b_n)$ segundo a expressão:

$$\sqrt{(a_1 - b_1)^2 + (a_1 - b_1)^2 + ... + (a_n - b_n)^2}$$

Árvore de Decisão

Com a utilização de árvores de decisão em classificação, pretende-se obter um modelo que, com base no conjunto anotado, defina um conjunto de regras para chegar à previsão da classe de uma nova instância. A analogia do modelo com uma árvore é que, começando no tronco, chega-se a uma bifurcação, que contém uma condição aplicada a uma das features, cujo valor presente na instância decide o ramo a seguir. Após um determinado número de bifurcações, chega-se à folha: a classe prevista. Uma das grandes vantagens deste modelo é que o conjunto de regras pode ser visualizado e uma das desvantagens é que podem ser gerados modelos demasiado complexos que se adaptam em demasia ao conjunto de treino

(*overfitting*), perdendo-se a capacidade de generalização.

A implementação neste trabalho usa o algoritmo CART (*Classification and Regression Trees*, caracterizado por construir árvores binárias, ou seja, um ramo sempre se divide em dois, nunca em mais. É compatível com valores numéricos e categóricos. Para além disso, é resiliente a valores perdidos, a outliers, e ainda diminui a complexidade da árvore. O critério usado para medir a qualidade da bifurcação foi o *índice Gini*.

2.3.3 Avaliação do Desempenho da Classificação

Aqui estão listadas as medidas de desempenho utilizadas no trabalho para avaliar a classificação, nomeadamente, as versões "*macro*" e "*weighted*" (pesada), bem como os cálculos necessários que as compõem.

A versão mais simples, "*micro*", contabiliza de forma global o número de verdadeiros-positivos, falsos-negativos e falsos-positivos. A versão "*macro*" calcula medidas para cada classe e obtém a sua média (simples), ignorando o eventual desbalanceamento. Já a versão "*weighted*" obtém um "peso" para cada classe a partir do seu número de instâncias corretas, tendo em conta o desbalanceamento.

Notação	**Descrição**
y	Conjunto dos pares $(real, previsto)$
$\hat{y}$	Conjunto dos pares $(real, previsto)$ que correspondem
L	Conjunto de *labels* (previsto)
y_l	Subconjunto de pares $(real, previsto)$ com $previsto = l$
$\hat{y}_l$	Subconjunto de pares $(real, previsto = real)$ com $previsto = l$
$P(A, B)$	$\frac{\lvert A \cap B \rvert}{\lvert A \rvert}$ para dois conjuntos A e B. Caso $\lvert A \rvert$, $P = 0$
$R(A, B)$	$\frac{\lvert A \cap B \rvert}{\lvert B \rvert}$. Caso $B = 0$, $R = 0$
$F_1(A, B)$	$2\frac{P(A,B) \times R(A,B)}{P(A,B)+R(A,B)}$

Tabela 2.3: Notação usada na tabela 2.4.

Média	**Precisão**	**Recall**	**F1**
"micro"	$P(y, \hat{y})$	$R(y, \hat{y})$	$F_1(y, \hat{y})$
"macro"	$\frac{1}{\lvert L \rvert} \sum_{l \in L} P(y_l, \hat{y}_l)$	$\frac{1}{\lvert L \rvert} \sum_{l \in L} R(y_l, \hat{y}_l)$	$\frac{1}{\lvert L \rvert} \sum_{l \in L} F_1(y_l, \hat{y}_l)$
"weighted"	$\frac{1}{\sum_{l \in L} \lvert \hat{y}_l \rvert} \sum_{l \in L} \lvert \hat{y}_l \rvert P(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L} \lvert \hat{y}_l \rvert} \sum_{l \in L} \lvert \hat{y}_l \rvert R(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L} \lvert \hat{y}_l \rvert} \sum_{l \in L} \lvert \hat{y}_l \rvert F_1(y_l, \hat{y}_l)$

Tabela 2.4: Adaptação das métricas *Precisão*, *Recall* e *F1* ao problema multi-*label*. Adaptado do site da biblioteca `Scikit-Learn`.

Adicionalmente, calcula-se *Accuracy* da mesma forma habitual:

$$Accuracy = \frac{\text{nº de Previsões Correctas}}{\text{nº Total de Previsões}}$$

Neste trabalho, as medidas de avaliação de desempenho serão obtidas em validação cruzada: ***10-fold cross-validation***. Na falta de um conjunto de teste, o conjunto de treino será dividido em 10 partes. Depois, o modelo de classificador será treinado com 9 dessas partes e testado na sobrante. O processo irá repetir-se mais 9 vezes, sempre se

treinando numa combinação única de 9 partes e testando numa parte diferente. Retiram-se as medidas de avaliação em cada uma das 10 iterações. No final, é calculada a média dos 10 valores e o seu desvio-padrão. Comparativamente ao testar sobre o próprio conjunto total de treino, a validação cruzada é resiliente a *overfitting*. Com base nas medidas é escolhido o modelo de classificação (algoritmo) e treinado sobre o conjunto total de treino.

2.3.4 Active Learning

Podemos definir *Active Learning* (Settles, 2009) como o processo de selecionar, entre um conjunto não anotado de instâncias, aquela cuja anotação é a de mais alta '*utilidade*' para o conjunto de treino atual. Ou seja, trata-se de uma metodologia de seleção de dados para anotação de forma não-aleatória, com vista em melhorar o conjunto de dados anotado da forma mais eficiente possível.

Existe uma variedade de critérios de '*utilidade*'. Neste trabalho é utilizada uma implementação da biblioteca **modAL**[4], que usa o critério de incerteza, selecionando a classificação menos confiante (Lewis and Catlett, 1994).

Sequência do Processo:

1. Obter conjunto de treino inicial.
2. Treinar classificador.
3. Aplicar classificador aos dados não anotados.
4. Selecionar a instância com maior incerteza na classificação.
5. Anotar a instância e adicionar ao conjunto de treino.
6. Repetir o processo desde o passo 2.

2.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) é uma técnica de análise de tópicos (*topic modeling*) desenvolvida por Blei et al. (2003). A ideia básica é de que cada documento pode ser descrito por uma distribuição de tópicos (assuntos) e que cada tópico pode ser descrito por uma distribuição de palavras. A técnica exerce uma tarefa semelhante a *clustering*, onde cada observação pode pertencer a mais do que um *cluster*. É exigida a predefinição do número de tópicos a ser extraído, embora existam algoritmos para estimar o número ideal. Para cada tópico é obtida uma sucessão dos termos presente no documento, cada um associado à probabilidade de estar relacionado com o tópico.

A implementação utilizada neste trabalho é da autoria de Řehůřek and Sojka (2010).

2.5 Trabalhos Relacionados

A informação partilhada pelas pessoas, de forma aberta, nas redes sociais, é de inegável valor para aplicações no domínio do marketing, mas também para outros, tal como a

[4] `https://modal-python.readthedocs.io/en/latest/`

investigação em saúde. Não é surpresa que, em anos recentes, vários estudos relacionados a saúde tenham apostado na aplicação de técnicas de mineração de texto e processamento de linguagem natural para analisar o que as pessoas dizem relacionado ao seu estado de saúde, tanto para confirmar conhecimento existente, como para encontrar formas alternativas de calcular métricas, acompanhamento do doente ou até alcançar novas descobertas que possam contribuir para a melhoria de tratamentos.

Para estudar o sentimento expresso em publicações de fóruns de saúde online, Bobicev and Sokolova (2018) aplicaram um classificador de texto. Anotaram 160 publicações para construir o conjunto de treino, em duas categorias: sentimento e problema de saúde. Na primeira, uma publicação podia ser anotada como "Encorajamento", "Confusão" e "Factos", sendo permitido não ser anotada ou conter várias opções.

A investigação recente tem recorrido principalmente ao Twitter, uma plataforma aberta de *microblogging* com uma API de acesso acessível.

O Twitter tem sido usado em vários estudos, dos quais são destacados alguns exemplos.

Toda a rede social tem a sua própria dinâmica, baseada em tráfego, tendências, popularidade, o que favorece a partilha de publicações de certas contas. Seria útil perceber que contas são tidas como referências principais na rede, principalmente em tópicos críticos: a maioria dos tweets partilhados sobre doenças são publicados por contas de confiança? Se não, podemos estar a sofrer uma epidemia de desinformação. Para chegar a uma conclusão para este problema, precisamos primeiro de entender a estrutura geral e a diversidade de entidades presentes no campo da discussão online. Isto é o que Beguerisse-Díaz et al. (2017) tentou fazer.

Sinnenberg et al. (2016) encontrou no Twitter uma ferramenta promissora para servir como fonte de dados para a investigação da doença cardiovascular, descobrindo pontos interessantes tal como o facto do termo "diabetes" aparecer mais vezes em tweets relacionados à doença em comparação com "insuficiência cardíaca". Detectou também uma maior prevalência em mulheres.

Também foi usado para detectar o mau entendimento ou má utilização de medicamentos (Scanfeld et al., 2010) tal como os seus efeitos adversos (O'Connor et al., 2014), bem como a monitorização de picos de gripe e outras epidemias em tempo real (Aramaki et al., 2011, Santos and Matos, 2014). Ainda sobre a gripe, Flekova et al. (2018) analisam tweets escritos antes, durante e após os utilizadores reportarem sintomas, tendo explorado os conteúdos emocionais, cognitivos e estruturais.

O trabalho de Yin et al. (2015) partilha mais semelhanças com este. Com o objectivo de desenvolver uma *framework* para detectar referências ao estado de saúde pessoal no Twitter, de forma automática e escalável, os autores começaram por recolher tweets via Streaming API. Após aquisição, aplicaram um filtro de palavras-chave para direcionar cada tweet a um problema de saúde específico. Extraíram um subconjunto de 2000 tweets focados em quatro problemas (cancro, depressão, hipertensão, leucemia) e o submeteram na plataforma Amazon Mechanical Turk[5], para ser anotado manualmente recorrendo a crowdsourcing. O primeiro passo foi perceber se o tweet mencionava efetivamente um estado de saúde pessoal. No caso afirmativo, os tweets foram também anotados de acordo com o doente, se corresponde ao autor do tweet, a um familiar ou amigo, ou a outra pessoa. No caso negativo, os tweets obtiveram um dos seguintes rótulos: metáfora, ponto de vista, preocupação ou nenhuma das anteriores. Foi treinado um classificador Naive Bayes multinomial para identificar as referências ao estado de saúde pessoal no conjunto

[5] `https://www.mturk.com/`

anotado, atingindo uma precisão de 0.77 quando aplicado a um corpus de 34 condições de saúde distintas. Foram consideradas as seguintes *features*: nomes, verbos, pronomes, dependências sintáticas, pontuação, emojis, URLs, hashtags e referêcnias a utilizadores. Foram retiradas várias conclusões para diferentes doenças, incluindo a probabilidade de um utilizador revelar a sua condição de saúde ou de outra pessoa. A conclusão que mais se destaca é a possibilidade de criar um classificador com base num pequeno conjunto de doenças que pode ser aplicado num conjunto de tweets que mencionam doenças não contempladas no treino. Para tal, os autores fizeram algum pré-processamento, incluindo substituir os nomes das doenças por um termo comum.

Foi possível notar que, com poucas exceções para o Português como Santos and Matos (2014), a maioria dos estudos se destinam à língua inglesa, não contemplando as pessoas que comunicam noutras línguas.

Capítulo 3

Definição de Objetivos

Havendo já interesse no reconhecimento de emoções, foi averiguada a existência de um grupo de doenças com relação particular ao estado emocional, de forma a reunir sinergicamente as competências que são próprias do ramo de informática com uma aplicação em saúde, indo ao encontro daquilo que é proposto no mestrado integrado: desenvolver competências de engenharia e aplicá-las na investigação e desenvolvimento de soluções biomédicas. Foi tomado o conhecimento das chamadas *psicodermatoses* a partir de uma entrevista dada em televisão em Maio de 2018[1], pelo vice-presidente da Sociedade Portuguesa de Dermatologia e Venereologia, António Miguel Peres Correia, a propósito da "Reunião da Primavera", um dos dois encontros realizados anualmente. As *psicodermatoses* foram aqui definidas como"*alterações da pele que são relacionadas com as perturbações emocionais que todos nós, de modo mais ou menos profundo ao longo da nossa vida, acabamos por nos defrontar*". O Professor Orientador conseguiu prover uma reunião com duas profissionais, médicas dermatologistas, do serviço de dermatologia dos Hospitais da Universidade de Coimbra: a Doutora Maria Margarida Martins Gonçalo e a Doutora Bárbara Roque Ferreira. Fomos aí alertados de que existem várias categorias de *psicodermatoses*, sendo do interesse o estudo da categoria mais consensual: os distúrbios psicofisiológicos.

As dermatoses nesta categoria, são doenças com origem fisiológica, i.e., não têm a sua origem num distúrbio psicológico como é o caso, por exemplo, da *tricotilomania*, uma doença caracterizada pelo vício de arrancar pelos e cabelos, despertado por um distúrbio psicológico. Informação mais detalhada pode ser encontrada em Duncan and Koo (2018). A relação com o estado emocional dá-se pelo agravamento da sintomatologia em condições de tensão e ansiedade. Para aumentar o sucesso do tratamento ou evitar a rápida expansão da doença, torna-se necessário normalizar o estado emocional, sendo de grande benefício complementar com acompanhamento psicológico. Seguem alguns exemplos de tweets:

> Desenvolvi uma psoríase por causa da depressão e do estresse e só depois disto procurei terapia. Faz 3 anos q ñ tenho nd

> To muito feliz que mês que vem recebo alta do meu tratamento de psoríase. Se você tem ansiedade ou depressão, procura um especialista ou alguém que você confie, não sofra sozinho!❤

[1] https://sicnoticias.pt/programas/edicaodamanha/2018-05-23-Dermatologistas-debatem-atualidade-das-doencas-da-pele-em-Portugal-esta-sexta-feira-e-sabado

Em reunião, foi definido o grupo de dermatoses que seria o foco deste estudo: **Psoríase**, **Urticária**, **Dermatite Atópica (também chamada de Eczema Atópico)** e **Vitiligo**. Foi ainda sugerida a análise de mais duas doenças muito comuns: **Diabetes**, para efeitos de controlo; e **Asma**, por estar relacionada com a Dermatite Atópica[2].

Este trabalho é de carácter altamente exploratório e não tem como objectivo fundamental o desenvolvimento de uma ferramenta para aplicação, mas sim a realização de um estudo para averiguar até que ponto é possível obter informação de interesse sobre estas doenças nas redes sociais.

O objetivo principal passa por perceber como as pessoas se expressam nas redes sociais em relação à sua doença, concretamente no Twitter [3], por, entre as várias redes sociais em que utilizadores partilham as suas experiências, se tratar da plataforma mais acessível à coleta de dados. Pretende-se perceber a expressão dos doentes em duas naturezas principais: o vocabulário utilizado e as emoções presentes.

De forma objetiva:

1. Obter indicadores de popularidade de cada doença, bem como a sua evolução temporal.

2. Adquirir o vocabulário associado às doenças estudadas, neste caso, através da análise de tópicos (em inglês, *topic modeling*), o que também pode ajudar a interpretar em que contextos as doenças são mencionadas.

3. Calcular o espectro emocional associado a cada doença.

4. Realizar uma análise comparativa dos espectros emocionais entre doenças.

5. Distribuir o conjunto de tweets ao longo do tempo, agregados por tipo de emoção.

6. No decorrer deste capítulo, surgirá a necessidade de completar lista de objetivos: Identificar quando o autor do tweet é realmente doente, através de métodos de classificação.

Acreditamos que, indiretamente, os resultados deste estudo podem contribuir para um melhor acompanhamento das doenças alvo, através de uma maior sensibilização para a correlação do estado emocional e da expressão patológica, ou para o desenvolvimento de ferramentas de monitorização do estado de saúde pública nas redes sociais compreendendo os períodos de maior ou menor popularidade. Poderão mesmo sugerir o grau de necessidade de fármacos ou a criação de meios online para a verificação de factos, semelhantes à verificação de fake news, no que toca a blogs não especializados de saúde, que podem sugerir tratamento ou automedicação nociva aos doentes. Ainda que não seja o objetivo principal, do ponto de vista exclusivamente tecnológico, o trabalho tornou mais claras as dificuldades em filtrar e caracterizar a informação de interesse, entre tudo o que é partilhado nas redes sociais.

[2]**Nota:** O termo *'dermatose'* é uma designação genérica para doenças de pele, não devendo ser confundida com *'dermatite'* que é utilizada para designar uma inflamação da pele. Dermatites são dermatoses, mas nem todas as dermatoses são dermatites.

[3]`https://twitter.com/`

3.1 Obtenção do conjunto de dados

O objeto de estudo neste trabalho consiste num conjunto de tweets que mencionam as doenças já referidas. A recolha de tweets foi efetuada várias vezes, porque em alguns momentos se verificou total ausência de tweets dentro de determinado intervalo de tempo, evidenciando o erro na obtenção dos IDs ou na aquisição. Inicialmente o conjunto abrangia os tweets desde 2008 (início da popularização da plataforma), até Junho de 2018, mas uma vez que foi necessário repetir o processo, acabaram por ser também adquiridos os tweets correspondentes à segunda metade de 2018. Para evitar problemas na obtenção dos dados, as pesquisas foram redefinidas para compreender intervalos de tempo mais curtos, sendo feita uma pesquisa por metade de cada mês. Para uma melhor comparação dos resultados com os capítulos seguintes, as experiências neste capítulo, que inicialmente não incluíam os dados da segunda metade de 2018, foram repetidas com a totalidade dos dados.

A obtenção dos tweets foi dividida em três momentos: a pesquisa dos tweets; a extração dos IDs (número único que identifica o tweet); e a recolha de tweets através da API do Twitter, assim que os IDs são submetidos.

Fez-se uso da API disponibilizada pelo Twitter para obter uma grande quantidade de tweets de forma automática, o que não seria possível manualmente, já que se tratam de 1 milhão de exemplares. Para obter acesso à API é necessário seguir um conjunto de passos:

1. Criar conta de utilizador comum[4].
2. Criar conta como desenvolvedor[5].
3. Criar uma *App*, sendo necessário justificar e esperar aprovação por parte do Twitter.
4. Uma vez aprovada, são disponibilizadas chaves únicas para a utilização da API.
5. É de extrema utilidade recorrer a uma biblioteca (SDK) destinada à linguagem de programação na qual se pretende trabalhar, que simplifique a utilização da API do Twitter. Neste caso, para a linguagem Java, foi utilizada a biblioteca *Twitter4J* [6].

A API oferece um conjunto variado de opções para pesquisa e obtenção de tweets. Uma das opções é realizar a pesquisa diretamente através da API, com a desvantagem de apenas apresentar resultados com máximo de 7 dias da sua publicação. Para além disso, nem todos os tweets são apresentados, apenas os considerados relevantes pela plataforma. Esta modalidade denomina-se *Standard Search API* [7].

Sendo assim, foi necessário recorrer a uma abordagem alternativa para obter tweets com mais de uma semana. A pesquisa através da página comum do Twitter permite obter tweets desde a sua fundação e sem filtro de relevância se for selecionada a opção '*Últimas*' em vez da opção-padrão '*Destaques*', que exclui a grande maioria. A opção '*Últimas*' ainda dispõe os tweets por ordem cronológica (inversa: mais recentes primeiro). Cada tweet é identificado por um ID, que uma vez submetido à API do Twitter é depois transferido desde que, entretanto, não tenha sido eliminado da plataforma, protegendo os utilizadores. Depois de realizar a pesquisa na página comum, é possível extrair os ID's dos tweets a partir do código `html`.

[4] `https://twitter.com/i/flow/signup`
[5] `https://developer.twitter.com/`
[6] `http://twitter4j.org`
[7] `https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html`

Segue-se o exemplo de uma pesquisa:

urticária -filter:retweets -filter:replies lang:pt since:2018-01-01 until:2018-01-15

No início é colocado o nome da doença. A pesquisa do Twitter é resiliente a diferentes utilizações de letras maiúsculas e minúsculas, de acentos, e até a erros ortográficos. Os termos utilizados foram: *psoríase*, *urticária*, *dermatite atópica*, *eczema atópico*, *vitiligo*, *diabetes* e *asma*. No final, os conjuntos de *dermatite atópica* e *eczema atópico* foram reunidos, removendo repetições. São adicionados ainda cinco modificadores à *query* de pesquisa:

- **-filter:retweets** o sinal '**-**' permite excluir os tweets identificados como *retweets* (partilhas realizadas dentro do Twitter, não contemplando as partilhas feitas a partir de páginas externas como sites de notícias);
- **-filter:replies** permite excluir tweets que consistem em respostas a um outro tweet;
- **lang:pt** restringe a pesquisa a resultados apenas escritos na língua portuguesa. Alguns tweets em língua espanhola resistem a este modificador;
- **since: e until:** estes modificadores permitem definir o período de publicação dos tweets a obter.

No início do trabalho, o ambiente em linguagem de programação Java foi adotado para todas as tarefas. Por uma questão de facilidade na utilização de algoritmos de classificação (nos capítulos seguintes), houve uma migração para o ambiente em Python 3. No entanto, duas tarefas foram conservadas em ambiente Java: a aquisição de tweets e a utilização de conjunto de ferramentas para o processamento base da linguagem, nomeadamente, a lematização, através das ferramentas incluídas no NLPPort (Rodrigues et al., 2018b).

Os tweets consistem em textos informais, onde os erros ortográficos são bastante abundantes, bem como o uso irregular de letra maiúscula e minúscula. Torna-se necessário processar o texto de forma a tornar possível que os seus elementos sejam comparáveis.

3.2 Pré-processamento do texto

Alguns tweets obtidos não mencionam qualquer doença, mas a sua aquisição justifica-se pela presença do nome de uma doença no nome do utilizador como, por exemplo, os utilizadores @*Asma_G95* e @*diabetes*. Algumas contas de organizações também referem a doença no seu nome de utilizador, mas nem sempre tratam sobre a doença no tweet.

A estratégia para resolver este problema passou por identificar referências da doença nos tweets e descartar aqueles que não contêm nenhuma. Os passos seguintes foram repetidos para cada doença.

1. Foi gerada um cópia de cada tweet em letras minúsculas;
2. A cópia foi inspecionada pela presença do nome da doença, com e sem acentos, e no caso de estar presente, foi substituído pelo termo '`DISEASEREF`';
3. Os tweets sem a presença de '`DISEASEREF`' foram agrupados;

4. Os tweets presentes neste grupo foram inspecionados e os conjuntos de termos compatíveis com a doença foram sendo anotados, repetido-se o passo 3 várias vezes, reduzindo o número de tweets a serem inspecionados muito rapidamente.

5. O passo 4 também foi agilizado, obtendo-se para cada tweet do grupo com ausência do termo '`DISEASEREF`', até três termos presentes, cuja sequência de carateres era mais semelhante ao nome das doenças alvo. Para tal, foi utilizando o método '`get_close_matches`' da biblioteca **difflib**[8], que implementa o algoritmo de Ratcliff-Obershelp [9]. Este processo resultou numa lista sem repetições, onde cada resultado foi inspecionado e adicionado à lista de variantes se, de facto, se tratasse de uma referência à doença. A lista final pode ser consultada na tabela 3.1.

Para o caso de dermatite atópica, também são consideradas as variantes de "*eczema atópico*". Menções semelhantes a "`#Dermatite #Atópica`" são convertidas em "`#DISEASEREF`" e posteriormente em "`HASHTAGREF_DISEASEREF`", de forma a conservar a informação que estamos na presença de uma `hashtag`, pois o algoritmo de *tokenização* remove o símbolo '`#`'.

Doença	**Variantes**
Psoríase	psoríase, psoriase, psoriáse, psoriática, psôriase, psorîase, psóríase, psóriase
Urticária	urticária, urticaria, úrticária, úrticaria, urtícaria, urticària, urticaría, urticariá
Dermatite Atópica	dermatite atópica, dermatite atopica, dermatiteatopica, dermatiteatópica, dermatite #atopica, dermatite #atópica, (dermatite) atópica, dermatite-atópica, dermatite atôpica, dermatite átopica, dermatíte atópica, dermatite..... atópica, dermatite átopica, dermatite atópica, dermatite 'atópica', dermátite atópica, eczema atópico, eczema atopico, eczemaatopico, eczemaatópico
Vitiligo	vitiligo, vitíligo, vitilígo, vítiligo
Diabetes	diabetes, díabetes, diàbetes, diábetes, diabétes, diabétés, diabètes, diabêtes, diabétés, diabeticos, diabéticos, diabéticas, diabtéticos, diabetico, diabete
Asma	asma, ásma, àsma, âsma, asmática, asmá, asmã

Tabela 3.1: Variantes das referências às doenças, reduzidas a letra minúscula.

# Tweets	**Psoríase**	**Urticária**	**Dermatite Atópica**	**Vitiligo**	**Diabetes**	**Asma**	**Total**
Obtidos	22.418	21.960	4.466	55.734	579.925	421.438	1.105.941
Sem menção	67	11	52	10	6923	3448	10511
Preservados	22.351	21.949	4.414	55.724	573.002	417.990	1.095.430

Tabela 3.2: Número total de tweets obtidos por doença e número de tweets que não a mencionam.

Como se pode observar na tabela 3.2, o número de tweets rejeitados é insignificante no caso das dermatoses, embora se destaque uma quantidade superior (em proporção) no caso da dermatite atópica, provavelmente associada ao facto de incluir duas palavras, o que torna o controlo das variantes mais difícil, até para o Twitter. No caso de diabetes e asma, o número é bastante expressivo, mas a proporção continua a ser insignificante. Daqui se sugere que a aplicação deste procedimento poderá nem ter grande impacto nos resultados e, eventualmente, não terá de ser realizado para todos os casos.

Ao longo deste trabalho, a necessidade de modificar alguns termos foi-se tornando evidente à medida que se implementavam algoritmos para vectorizar os tweets: endereços URLs, emoticons e emojis eram desfeitos, surgindo na lista de features, obtidas da tokenização automática para bag-of-words, pedaços de cada tipo de elemento citado. A partir

[8] `https://docs.python.org/3/library/difflib.html`

[9] `https://collaboration.cmc.ec.gc.ca/science/rpn/biblio/ddj/Website/articles/DDJ/1988/8807/8807c/8807c.htm`

do recurso a expressões regulares, foram identificados os endereços URL e substituídos pelo termo '`URLREF`'. Quanto a emojis e emoticons, constituem elementos evidentemente carregados de informação emocional, logo não podiam ser descartados ou meramente substituídos por um único termo. Recorrendo a uma biblioteca em Python, denominada simplesmente por **emoji**[10], os emojis foram localizados nos tweets, convertidos à sua descrição textual (`alias`) e anexados ao prefixo '`EMOJI_`'. Por exemplo:

😠 → `EMOJI_angry_face`

Desta forma, todo o termo '`EMOJI_alias`' foi conservado após o processo de tokenização presente nas implementações da biblioteca **Scikit-Learn**[11].

Os emoticons[12] foram também substituídos, com alguns cuidados como se descreve mais à frente. Para identificar estes elementos, os tweets foram inspecionados pelos exemplares contidos numa lista disponibilizada por Rodrigues et al. (2018a), que contém também a sua descrição textual (`alias`). Alguns emoticons apresentam várias descrições textuais possíveis, então foi escolhida a primeira. À semelhança do caso dos emojis, foi anexado um prefixo:

`:(` → `EMOTICON_frown`

Foi também realizada a substituição das referências a utilizadores (@**user** → `USERREF`), hashtags (**#exemplo** → `HASHTAG_exemplo`), e o mais importante, a doenças (**diabetes** → `DISEASEREF`), tornando o sistema de classificação mais versátil, ainda mais necessário quando não se inclui no conjunto de treino todas as doenças envolvidas no estudo. O termo `DISEASEREF` estará presente em todos os tweets e deste modo os classificadores não absorverão informação discriminatória através dela, o que não ocorreria se os nomes das doenças fossem preservados. Para evitar isso, podia ser sugerido remover até a presença do termo `DISEASEREF`, mas perder-se-iam features importantes em n-gramas, por exemplo, o bigrama '`tenho DISEASEREF`' constitui um indício muito evidente para classificar o autor do tweet como doente, uma necessidade concluída no final do presente capítulo e descrita nos seguintes.

Para além da substituição de referências a doenças, hashtags, utilizadores, URL's e a emojis/emoticons, foi também obtida a versão lematizada de cada tweet recorrendo à ferramenta **LemPORT**, desenvolvida por Rodrigues et al. (2014).

Um cuidado muito importante deve ser tido em conta de forma a não distorcer a informação contida no tweet: a ordem do processamento.

- Substituir endereços `URLs` antes de `emoticons` e nome da doença: os endereços podem conter o nome da doença, bem como símbolos que reunidos constituem um emoticon (p.ex., em 'http://' temos o emoticon '`:/`').
- Substituir emoticons por suas descrições antes de hashtags e nomes de utilizador: alguns emoticons incluem o símbolo '`#`' e '`@`'.
- Substituir emoticons antes de reduzir o tweet a letra minúscula, de forma a não

[10] `https://pypi.org/project/emoji/`
[11] `https://scikit-learn.org/`
[12] Emoticons: agrupamento de caracteres para formar expressões faciais.

converter '`=D`' para '`=d`', por exemplo.

- Na substituição dos emojis pelos aliases, substituir *hífens* ('`-`') por *underscore* ('_'), de forma a não ser separável na tokenização. Os aliases dos emoticons listados não contêm *hífens*.

3.3 Primeiras Análises aos Dados Recolhidos

Esta secção apresenta os resultados de um conjunto inicial de experiências realizadas sobre o conjunto dados recolhidos, do pré-processamento do texto, nomeadamente:

- **Análise de Tópicos**: é aplicada uma técnica de *topic modeling* ao conjunto de tweets das quatro dermatoses, de forma a facilitar a identificação dos assuntos mais relevantes.
- **Análise do Espectro de Emoções**: recorrendo a léxicos de emoções baseados em palavras e emojis/emoticons, são obtidos os gráficos das emoções detetadas para cada doença.
- **Distribuição no tempo**: para cada doença, os tweets são agrupados pelo mês de publicação e expostos graficamente para analisar a popularidade ao longo do tempo.
- **Nacionalidade**: é feita ainda uma análise sobre até que ponto a informação da origem geográfica disponível é suficiente para restringir os resultados a Portugal e/ou Brasil.

Para além de permitirem uma primeira análise aos dados, os resultados obtidos motivam o restante trabalho realizado e ajudam a redefinir os seus objetivos.

3.3.1 Análise de Tópicos

Com o objetivo de analisar a diversidade de assuntos abordados, a técnica de LDA, muito utilizada para topic modelling, foi aplicada ao conjunto de tweets que mencionam pelo menos uma das quatro dermatoses. Apesar de ter sido experimentada com diferentes números de tópicos, optámos por realizar a análise sobre o resultado para 20 tópicos. Por mais abstrata que essa decisão possa ser, com base nas experiências realizadas, acreditamos ser um número suficiente para abranger uma boa diversidade. Os tópicos resultantes são apresentados na tabela 3.3 e discutidos de seguida.

O tópico 0 sugere a presença de tweets informativos sobre várias psicodermatoses (psoríase, vitiligo, dermatite atópica) como sendo doenças 'crónicas', sem 'cura', e não 'contagiosas'.

O tópico 1 traduz um evento de publicidade abusiva no Twitter, por parte de uma clínica situada em Sete Lagoas, no Brasil, que oferece sessões de 'terapia de casal', tratamento contra a depressão e até contra lúpus e vitiligo, provavelmente pela via do acompanhamento psicológico para a normalização do estado emocional. Os números presentes pertencem ao seu contacto telefónico. Este episódio de publicidade abusiva pode ser evidentemente visualizado na figura 3.5, correspondente à distribuição temporal dos tweets que mencionam vitiligo, entre Junho de 2016 e Janeiro de 2017. Uma inspeção aos dados permitiu concluir que perto de 80% dos tweets neste intervalo pertencem a esta ação publicitária.

#	Tópico
0	URLREF psoriase pele doença vitiligo tratamento via dermatite_atopica não cura mais crônica saúde doenças contagiosa causa vida natural você podem
1	vitiligo URLREF sete lagoas terapia casais #setelagoas depressao 3773 😂 conheça 8771 9185 cientistas avaliacao lúpus desenvolver cria quebra emoções
2	URLREF vitiligo psoriase causas encontro psoríase grande dr outubro sobre municipal :/ dermatologia evento sociedade paulo cid câmara artigo mostra
3	vitiligo meu to eu parece tenho cara descascando parecendo nao minha rosto mano tô vcs eh sol acesse agora alimentos
4	vitiligo URLREF michael ele jackson modelo não branco ela filho mundo era doença mais sua cor cara moda gente prince
5	urticaria alergia veja coceira dermatite_atopica :(alergias dor harlow água alérgica pior winnie droga anti momento ajudam efeitos meio
6	minha urticaria psoriase eu não to cure mãe meu agora quero mais vida boa noite tô ela dermatite_atopica estou causa
7	sintomas tratamentos eczema médicos estudo pacientes pouco nova dermatite_atopica medicamentos mais cubanos tema estão seus aguento suas realiza palestra novas
8	psoriase URLREF dia mundial hoje sobre vídeo campanha nacional 29 gostei doença população conscientização portadores preconceito bom combate 10 vamos
9	eu não tenho urticaria dermatite_atopica psoriase vitiligo estou muito meu sou acho mais fico minha bem ela deus pessoa aí
10	URLREF psoriase vitiligo tratar mais você tratamento pessoas contra remédios crianças têm 5 mulheres dermatite_atopica livro risco remédio conhece entenda
11	acompanhamento toc URLREF vitiligo amigos post pesquisa cabelos jovem site primeiro portadora brancos espontânea portugal iniciativa facebook vários voz vira
12	nome sus povo paciente base causar especialista sai medicamento direito artrite passa pega consegue resultado levar ok cobre fique dieta
13	vitiligo el às forma favor dra nesta and junho corpo vezes maria comuns maltrata friends chegando pelas malhado próxima depoimento
14	urticaria me URLREF gente la dando pensar chega causa ouvir provoca tanta con vejo palavra el le produce ler pessoas
15	sobre saiba vitiligo fala mais dúvidas falando programa bem doenças matéria tv assunto estava algumas diagnóstico dermatologista entrevista hospital globo
16	vitiligo confira ela dela técnica naturais planta falou #saúde nosso 😍 cabelo cerrado esperança new <3 13 angioedema laser origem
17	vitiligo URLREF mancha maquiagem branca foto usava peruca dicas jackson mulher confirma on homem participe parece preta live autópsia limão
18	urticaria associação frio 😭 puta 6 combater colinérgica calor inverno sofrer escola tempo estado 30 7 neste top acabar ouviu
19	anos 1 vitiligo 2 3 manchas há brancas aos 4 dias brasil cuidados cerca umas apenas 10 tatuagem 8 desde

Tabela 3.3: 20 tópicos obtidos com LDA a partir de **todos os tweets** das quatro psicodermatoses.

O tópico 2 trata de um evento de sensibilização sobre dermatoses.

O tópico 3 sugere a discussão sobre influências, como a luz solar, sobre a condição de vitiligo, a ocorrência no rosto/cara, a influência dos alimentos e a menção a sintomas. No entanto, ao inspecionar o conjunto de dados, foi verificado que se trata de uma referência a vitiligo em termos de comparação: devido a queimadura solar e consequente renovação da pele, utilizadores expressam que 'parece' que estão com vitiligo.

O tópico 4 refere Michael Jackson e relaciona-se com as várias polémicas também tratadas nas redes sociais, sobre a causa da mudança da coloração da sua pele, se natural (vitiligo) ou artificial. Faz-se presente também 'Prince', não o artista, mas o filho de Michael Jackson que, possivelmente, também sofrerá de vitiligo.

O tópico 5 trata da urticária, também na sua forma aquagénica (alergia à água), mencionando um sintoma ('coceira') e emoção negativa (':('). Outros termos parecem constar como ruído, como 'dermatite atópica' ou a referência a Winnie Harlow, modelo famosa portadora de vitiligo.

O tópico 7 consiste possivelmente num lapso: refere 'médicos cubanos', muito publicitados em 2016 pela alegada descoberta de uma cura para vitiligo, mas aparece num tópico que inclui 'dermatite atópica' e 'eczema'. 'Eczema atópico' é uma designação equivalente

para dermatite atópica, no entanto, apenas 'eczema' pode ser aplicado a uma grande variedade de dermatoses.

O tópico 8 trata do dia mundial da psoríase, a 29 de Outubro, popular pelos eventos de sensibilização ('concientização') sobre a psoríase e o estigma social causado pela suspeita de ser contagiosa. Na figura 3.2 pode ser observada uma série de picos, igualmente espaçados, pelos meses de Outubro.

O tópico 10 sugere o reencaminhamento para uma página externa de artigos de informação ('URLREF','conhece','entenda') sobre as doenças, tratamentos, remédios e riscos.

O tópico 11 parece misturar assuntos distintos: 'acompanhamento' e 'toc' (Transtorno obsessivo-compulsivo) costumam ser também mencionados nos tweets de publicidade relatados no tópico 1. Os termos 'amigos', 'vitiligo', 'facebook' e 'portugal' podem remeter para um grupo nessa rede social. Já 'cabelos' 'brancos' pode remeter para a ocorrência de vitiligo no couro cabeludo, chamada de poliose.

O tópico 12 faz referência a 'sus', acrónimo para 'Sistema Único de Saúde', denominação do sistema público de saúde no Brasil. Este tópico é enigmático já que não menciona nenhuma das quatro dermatoses, fazendo referência a artrite. No site da Sociedade Brasileira de Dermatologia (SBD) [13] pode-se ler "*Segundo as especialistas, nem todos os portadores da psoríase precisam de tratamentos sistêmicos. A maioria tem lesões leves, tratáveis com remédios de uso tópico. Porém, 30% dos pacientes têm crises graves, que acometem o corpo todo, ardem, coçam e chegam a desenvolver a chamada "artrite psoriática", que inclui sintomas como dor, rigidez e inchaço nas articulações.*". Ficará assim justificada a razão de ocorrência de 'artrite' dentro do tópico, uma consequência não óbvia de se ter psoríase.

O tópico 14 aparenta ter alguma influência de tweets em língua espanhola ('la', 'con', 'el', 'produce'), mas o sentido é bastante óbvio: o uso figurado de 'urticária' para representar a irritação causada por certas pessoas ('gente').

O tópico 15 reflecte a publicidade feita a um 'programa' no canal de televisão ('tv') Globo, cujo assunto ('matéria') são condições de saúde ('doenças'), das quais se inclui vitiligo. Trata-se de uma 'entrevista' a um profissional 'dermatologista' onde o telespectador pode esclarecer as suas 'dúvidas', inclusivamente sobre o 'diagnóstico'.

O tópico 16 sugere a presença de tweets que recomendam um tratamento 'natural', baseado em plantas ('planta'), e de outros que sugerem uma 'técnica' 'laser'. Vitiligo e urticária são mencionados e sobre a segunda, é abordado o 'angioedema'. Segundo a Sociedade Brasileira de Dermatologistas [14], consiste num *"inchaço rápido, intenso e localizado, que atinge normalmente pálpebras, lábios, língua e garganta."*. O tópico inclui ainda um emoji e um emoticon, ambos de valência positiva.

O tópico 17 parece tratar do mesmo assunto noticiado em 9 de Fevereiro de 2010, no jornal online G1 (Globo) [15], intitulado *"Necrópsia confirma que Jackson usava peruca e tinha vitiligo"*. Neste tópico fazem-se incluir ainda 'dicas' de 'maquilhagem' para 'mulher', provavelmente para disfarçar 'manchas' 'brancas' de vitiligo. O 'limão' é vulgarmente recomendado como tratamento para vitiligo em blogs populares.

O tópico 18 expressa uma emoção negativa através de emoji (😭) e o calão ('puta'),

[13] `https://www.sbd.org.br/psoriasetemtratamento/noticias/informe-se/sus-pode-adotar-novo-tratamento-da-psoriase/`

[14] `https://www.sbd.org.br/dermatologia/pele/doencas-e-problemas/urticaria/73/`

[15] `http://g1.globo.com`

no contexto de subtipos de urticária, onde os sintomas são desencadeados em situações de mudança de temperatura. Este é o caso da variante 'colinérgica', caracterizada pela expressão de sintomas com o aumento de temperatura corporal ('calor'), e também da chamada urticária de 'frio' [16], que se agrava naturalmente durante o 'inverno'.

Do tópico 19 sobressai o termo 'tatuagem', pela partilha de vários artigos sobre se existe algum perigo em alguém que sofre de vitiligo fazer tatuagem.

O tópico 6, bem como o 9, apresentam-se bastante heterogéneos. O tópico 13 segue o mesmo registo, incluindo termos em língua inglesa ('and', 'friends') e talvez espanhola ('el').

Apreciação

Na tabela 3.3, podemos contar **9** tópicos contendo '`URLREF`'. Dado que este termo substitui qualquer endereço URL, sugere que uma grande quantidade de tweets contem uma ligação externa, o que sugere que o tweet não descreve propriamente o estado de saúde pessoal do seu autor, tendo em conta a observação meramente empírica da parte do autor do presente estudo, a partir do contacto com o conjunto de tweets. Da mesma forma, existem muitos tópicos que sugerem a presença de partilhas de notícias, eventos e curiosidades sobre personalidades populares. É por isso necessário recorrer a algum tipo de filtragem, resiliente à ocorrência de menções a doenças no sentido figurado (muito comum para urticária), que permita fazer uma análise mais voltada para tweets que realmente transmitem o estado de saúde pessoal do autor. Na tabela analisada contam-se **3** emojis e **3** emoticons. Será de esperar que, se fazendo uma filtragem adequada, ocorra um maior número de expressões emocionais. O tópico 6 é o único tópico que se apresenta compatível com tweets que mencionam realmente o estado de saúde pessoal. Eventualmente o número 9 também pode ser compatível, mas a presença do termo 'não' no início do tópico assinala a sua relevância, podendo representar tweets onde o autor nega a presença de uma dermatose, de forma semelhante ao tópico 3.

3.3.2 Emoções

Para se obter a emoção representativa de um tweet é necessário analisar o seu conteúdo. Tal como existem léxicos que associam definições ou informação morfológica às palavras de uma língua, existem léxicos que lhes associam informação relativa às emoções normalmente transmitidas. Dada a sua simplicidade, a exploração de léxicos deste tipo em análise de sentimentos ou reconhecimento de emoções acaba por ser bastante utilizada. Isto, se os léxicos do tipo desejado, efetivamente, existirem. Neste contexto, existem léxicos que associam emoções tanto a palavras como a emojis e emoticons, o que faz sentido dada a interligação entre muitos destes símbolos e a emoção que se pretende transmitir. A emoção pode ser representada através de uma simples palavra (modelos categóricos) ou através de valores de valência, ativação e outros.

No léxico LEED (Rodrigues et al., 2018a), são apresentados emojis com o mesmo `alias`, mas com pequenas diferenças nas imagens e nos valores de valência e ativação. Tal se deve ao facto da lista incluir variantes de emojis de várias plataformas (Android, iOS, Facebook, Emojipedia). Não incluindo as próprias do Twitter, foi calculada a média dos valores para o mesmo `alias`, e utilizada na análise com léxicos.

Para palavras, foi utilzado o léxico ANEW-PT, original em língua inglesa e adaptado para a portuguesa de forma controlada por Soares et al. (2012). Mais recentemente foi

[16] `https://www.sanfil.pt/urticaria-de-frio/`

publicado um léxico, NRC-VAD [17], que inclui um maior número de termos, mas cuja versão em português decorre de uma tradução automática. Por essa razão, apresentam-se os resultados com base no léxico ANEW-PT. O léxico ANEW-PT não inclui diferentes valores para a mesma palavra em inglês. No entanto, em alguns casos foi obtida a mesma tradução portuguesa para distintas palavras em inglês, o que obrigou a fazer a mesma ponderação entre valores como no caso dos emojis.

Na análise que recorre a léxicos torna-se necessário utilizar a versão lematizada dos tweets, porque é nessa forma que as palavras estão listadas no léxico.

Existem várias formas de conciliar as palavras e os emojis/emoticons para se obterem as coordenadas resultantes por tweets. A forma mais óbvia é, num só cálculo, contabilizar todos os elementos presentes e calcular a média dos valores de valência e ativação. Outra seria contabilizar apenas os emojis/emoticons presentes, e na sua ausência, contabilizar as palavras. Os resultados apresentados aqui decorrem de uma lógica diferente, definida após um conjunto preliminar de experiências:

1. Dada a forte ligação entre emojis/emoticons e emoções (veja-se, por exemplo Wood and Ruder, 2016), se existir um emoji/emoticon no final do tweet, os seus valores de valência e ativação são atribuídos ao tweet. Se vários elementos gráficos se encontrarem no final do tweet, é calculada a média dos seus valores.
2. Se não existir um emoji/emoticon no final, é calculada a média de todos os elementos gráficos presentes no tweet.
3. Só no caso de não existirem elementos gráficos presentes é que se calcula a média dos valores das palavras do tweet presentes no léxico ANEW-PT.
4. No caso de ausência de elementos incluídos nos léxicos, é atribuída a neutralidade de valência e ativação.

Os resultados de todas as lógicas aqui referidas são muito semelhantes. A lógica enumerada foi escolhida por apresentar uma melhoria residual de desempenho quando testada no conjunto anotado via crowdsourcing, que é tratado no capítulo seguinte. De qualquer maneira, o desempenho é claramente baixo, obtendo-se uma medida F1 de 0.27 para valência e 0.05 para ativação, quando testada sobre a totalidade do conjunto anotado, sem validação cruzada. Nas mesmas condições e sem o uso da melhor combinação de algoritmo e do tipo de features, foram obtidas através de classificação supervisionada, as medidas F1 de 0.95 e 0.94, para valência e ativação, respetivamente, o que levou mais à frente a rejeitar a abordagem baseada em léxicos.

A figura 3.1 apresenta o espectro emocional simplificado a três valores de valência e de ativação.

De seguida é apresentada a ordenação decrescente das percentagens dos grupos emocionais de valência negativa:

V-A+ (vermelho): psoríase (16.8%), asma (16.3%), dermatite atópica (14.1%), diabetes (11.8%), urticária (8.7%), vitiligo (5.1%).

V-A0 (laranja): nenhum conjunto alcançou 0.1% nesta categoria.

V-A- (amarelo): asma (3.63%), psoríase (2.25%), dermatite atópica (2.24%), urticária (2.21%), diabetes (1.76%), vitiligo (1.38%).

[17] `https://saifmohammad.com/WebPages/nrc-vad.html`

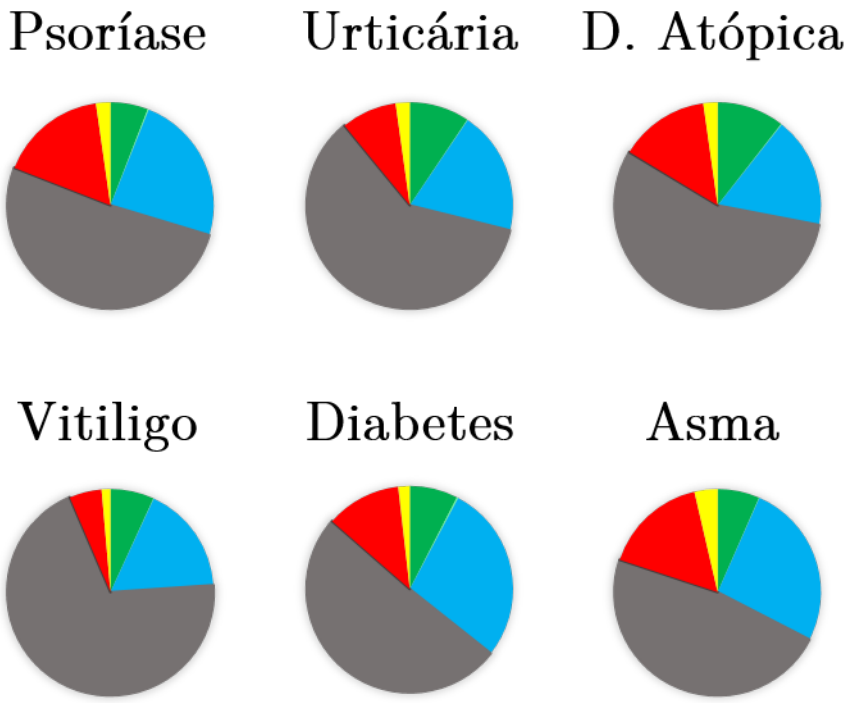

Figura 3.1: Espectro emocional do conjunto total de tweets respectivo a cada doença, por todos os quadrantes considerados e suas fronteiras. Resultados obtidos da análise de léxicos LEED e ANEW-PT.

Destaca-se a grande proporção de tweets de valência neutra (cinzento), e a maior proporção de tweets identificados com valência positiva (azul e verde) em relação aos tweets identificados com valência negativa (amarelo e vermelho).

A proporção de tweets associada às fronteiras (ativação média) dos quadrantes vermelho e amarelo (V-A0, laranja) e dos quadrantes azul e verde (V+A0, azul-esverdeado) não são percetíveis, porque esta simplificação não á adequada aos valores obtidos pelos léxicos: as regiões intermédias são caracterizadas por um só valor discreto e não um intervalo contínuo de valência/ativação.

Tendo em conta a análise de tópicos, a grande proporção de tweets neutros percetível nos espectros emocionais parece estar associada à presença de tweets que não mencionam o estado de saúde pessoal do autor. Assim se evidencia de forma especial nos conjuntos respetivos às duas doenças com maior proporção de tweets neutros: urticária, pelo uso conhecido da doença no sentido figurado; e vitiligo, por um grande número de publicações que comentam a mudança da coloração de pele de Michael Jackson, ou fazem referência a Winnie Harlow, uma modelo famosa que porta vitiligo.

3.3.3 Distribuição de tweets no tempo

Uma outra análise consiste em distribuir os tweets recolhidos e ao longo do tempo. As publicações estão agrupadas por meses listados no eixo horizontal, desde Janeiro de 2010 a Dezembro de 2018, e no eixo vertical medem-se o número de tweets publicados no mês correspondente.

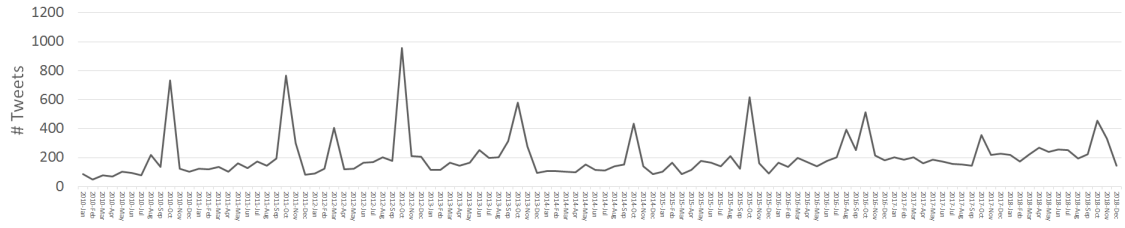

Figura 3.2: Distribuição temporal de tweets referentes a **Psoríase** entre 2010 e 2018.

Podemos ver na figura 3.2, a presença de picos igualmente espaçados, localizamos nos meses de Outubro. Tal se deve à comemoração do Dia Mundial da Psoríase, a 29 de Outubro, caracterizada pela publicação de informação, sensibilização e eventos por parte de organizações.

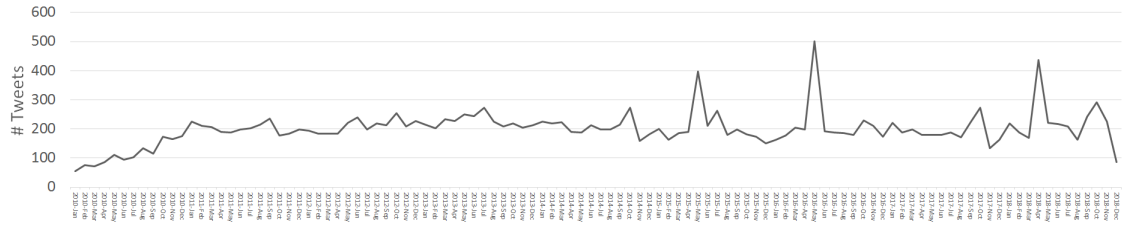

Figura 3.3: Distribuição temporal de tweets referentes a **Urticária** entre 2010 e 2018.

Na curva da figura 3.3 observam-se três picos maiores: Maio de 2015, Maio de 2016 e Abril de 2018. No dia 22 de Maio de 2015 foi publicada pelo grupo Globo a notícia:

"*Rinite é a alergia mais comum entre os brasileiros, seguida da urticária*"[18].

A partilha no Twitter está na causa deste pico. No dia 6 de Maio de 2016 foi imensamente partilhada uma notícia[19] que retrata a luta jurídica de uma mulher por obter comparticipação de um tratamento para *urticária crónica espontânea*, que custava aproximadamente R$ 5000 (≈ 1200 €, à data), por mês. Abril de 2018 foi caracterizado por vários eventos relacionados a urticária, como a partilha de endereços para um vídeo no Youtube, programas televisivos e campanhas de sensibilização sobre a doença.

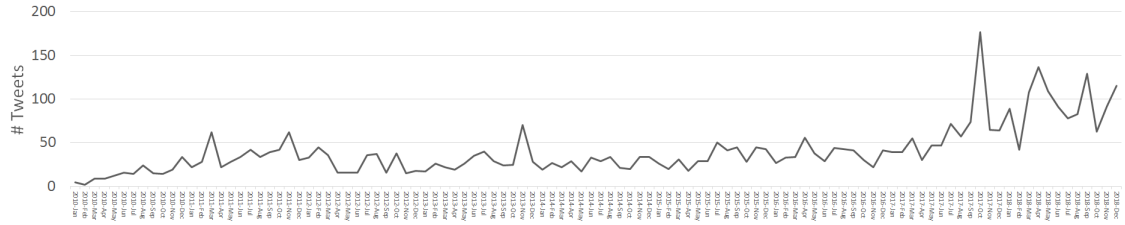

Figura 3.4: Distribuição temporal de tweets referentes a **Dermatite Atópica** entre 2010 e 2018.

Na figura 3.4 podemos notar que o maior pico ocorreu em Outubro de 2017. Neste mês, um programa televisivo denominado *"Bem Estar"*, publicitou um episódio que iria abordar dermatite atópica.

[18] `http://g1.globo.com/globo-reporter/noticia/2015/05/rinite-e-alergia-mais-comum-entre-os-brasileiros-seguida-da-urticaria.html`

[19] `https://t.co/L4bBvwOX8W`

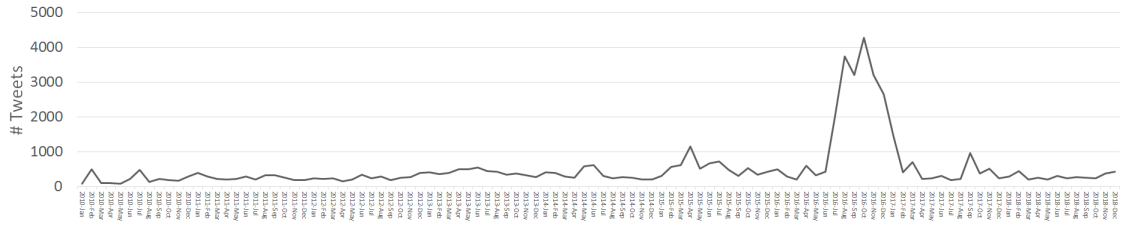

Figura 3.5: Distribuição temporal de tweets referentes a **Vitiligo** entre 2010 e 2018.

Na figura 3.5 é notável uma explosão de tweets entre os meses de Julho de 2016 e Janeiro de 2017. Trata-se de um episódio de publicidade abusiva, onde uma clínica, localizada em Sete Lagoas (Brasil), publicou uma grande quantidade de tweets, com pequenas variações, comunicando os serviços que disponibilizava de acompanhamento na depressão, terapia de casal, transtorno obsessivo-compulsivo, e até vitiligo.

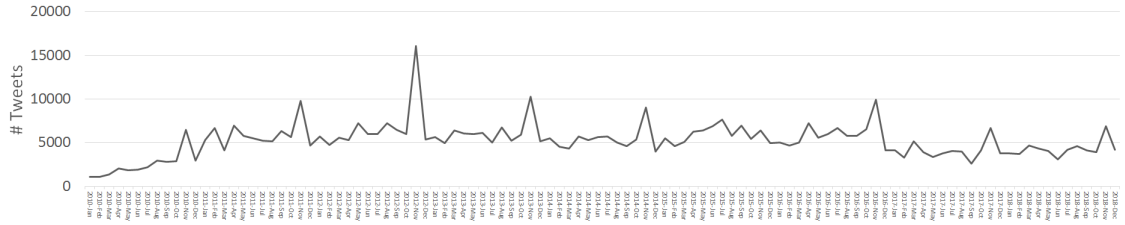

Figura 3.6: Distribuição temporal de tweets referentes a **Diabetes** entre 2010 e 2018.

A curva na figura 3.6 apresenta picos nos meses de Novembro, a propósito do Dia Mundial do Combate a Diabetes, a 14 desse mês.

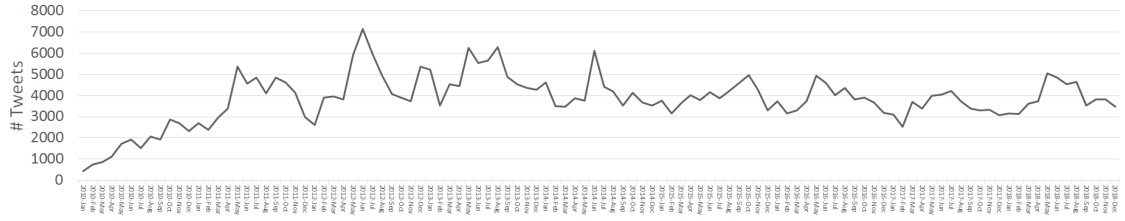

Figura 3.7: Distribuição temporal de tweets referentes a **Asma** entre 2010 e 2018.

Na figura 3.7 podemos observar até certo ponto, uma oscilação na curva com período de duração anual. A curva é bastante irregular, mas é possível notar esta tendência.

Na figura 3.8, estão presentes as curvas de distribuição temporal de todas as doenças. O único momento onde a curva de uma das dermatoses (vitiligo) se aproxima dos números de diabetes ou asma, é no episódio de publicidade abusiva.

3.3.4 Comparação Portugal-Brasil

Para além dos dados extraídos com o texto do tweet, como o ID, o nome de utilizador e a data de publicação, foram avaliados os atributos responsáveis por guardar informação geográfica, para averiguar a possibilidade de restringir o estudo aos tweets provenientes de Portugal.

Na biblioteca `Twitter4J`, o objeto no qual é guardada toda a informação do tweet obtido é denominado `Status`. Existem, para além das coordenadas de latitude e longitude,

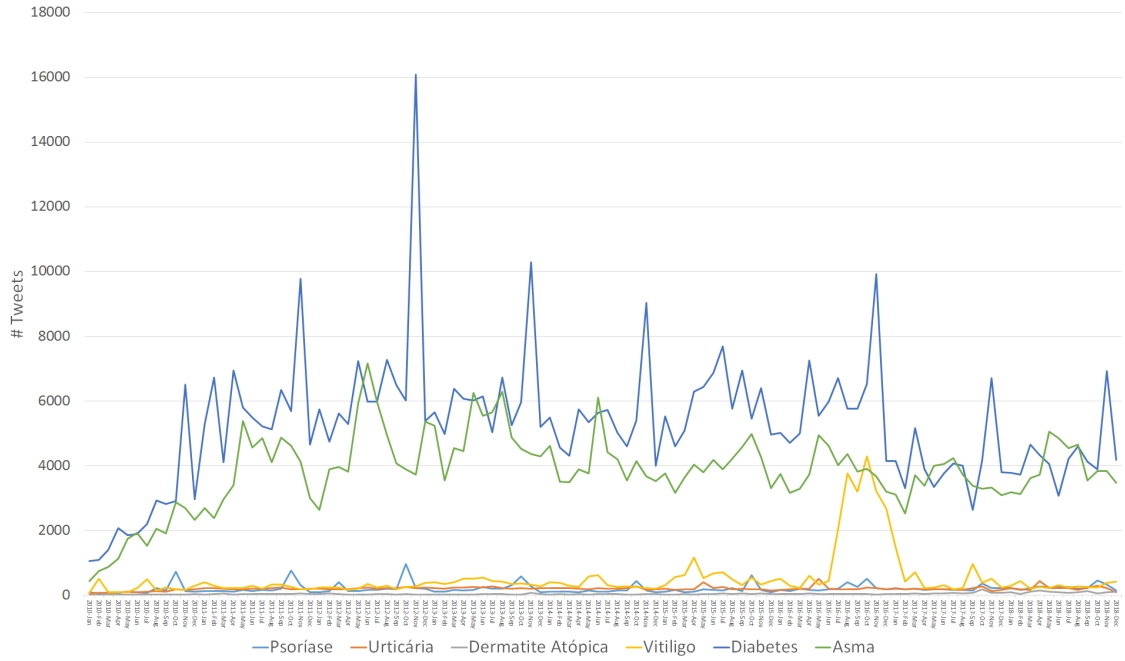

Figura 3.8: Comparação da distribuição temporal de tweets referentes às seis doenças abordadas entre 2010 e 2018.

muito ausentes porque a sua partilha exige a autorização expressa do utilizador, dois outros atributos que podem identificar a origem geográfica do tweet. Um deles pode ser acedido a partir do método `statusObject.getPlace().getCountryCode()` que devolve 'PT' no caso de Portugal e 'BR' no caso do Brasil. Outro atributo, pode ser obtido do método '`statusObject.getUser().getLocation()`', contendo algumas vezes o nome da cidade onde o utilizador reside, ou outra qualquer informação, já que este atributo é totalmente personalizável. Nenhum dos atributos mencionados está obrigatoriamente preenchido. Com base neles, tentou-se extrair a informação geográfica da sua origem. Sendo o atributo `CountryCode` mais fidedigno, foi inspecionado em primeiro lugar. O tweet que contivesse 'PT' era logo considerado português e aquele que contivesse 'BR' era logo considerado brasileiro. Aquele que contivesse este atributo preenchido, mas diferente de 'PT' e 'BR', foi catalogado como 'outro' e aquele que se apresentasse vazio seria sujeito à inspeção do atributo `Location`. Caso estivesse vazio, o tweet era catalogado como 'desconhecido'. Se incluísse, depois de ser reduzido a letra minúscula, a palavra 'portugal' sem mencionar 'brasil' ou 'brazil', era considerado português. Se mencionasse 'brasil' ou 'brazil' e não mencionasse 'portugal', era considerado brasileiro. Noutra qualquer opção era catalogado como 'indefinido'.

Acontece que muitos tweets contêm no atributo personalizável o nome da cidade do utilizador, mas não era possível deduzir diretamente o país de forma automática, já que, por exemplo, existem cidades com o mesmo nome em ambos os países, como é o caso de 'Coimbra', presente no estado de Minas Gerais ou na Barra do Piraí, no Rio de Janeiro. Com base nesta catalogação são apresentados os resultados nas tabelas 3.4 e 3.5.

# Tweets	**Psoríase**	**Urticária**	**Dermatite Atópica**	**Vitiligo**	**Diabetes**	**Asma**	**Todas**
Portugal	907	591	132	201	11.327	6.374	19.532
Brasil	7.454	5.024	1.540	20.753	171.221	120.140	326.132
Outro/indefinido	13.990	16.334	2.742	34.770	390.454	291.476	749.766

Tabela 3.4: Número de tweets provenientes de Portugal, Brasil e local indefinido.

% Tweets	Psoríase	Urticária	Dermatite Atópica	Vitiligo	Diabetes	Asma	Todas
Portugal	4.06	2.69	2.99	0.36	1.98	1.52	1.78
Brasil	33.35	22.89	34.89	37.24	29.88	28.74	29.77
Outro/indefinido	62.59	74.42	62.12	62.40	68.14	69.73	68.44

Tabela 3.5: Percentagem de tweets provenientes de Portugal, Brasil e local indefinido.

Após este cálculo, foram criadas as figuras de distribuição temporal só para Portugal. Com base nelas foi decidido não restringir os dados a estudar, já que o mecanismo de catalogação de país é muito falível, devido à abundante ausência de informação. Mais de metade dos tweets não constam como brasileiros, o que não é uma possibilidade admissível. De seguida, apresenta-se o gráfico da distribuição de tweets catalogados como portugueses, que mencionam dermatite atópica. Havendo meses sem contemplar tweets e o máximo nem superar 10, qualquer análise nunca poderá ser considerada representativa. As restantes distribuições podem ser consultadas em anexo.

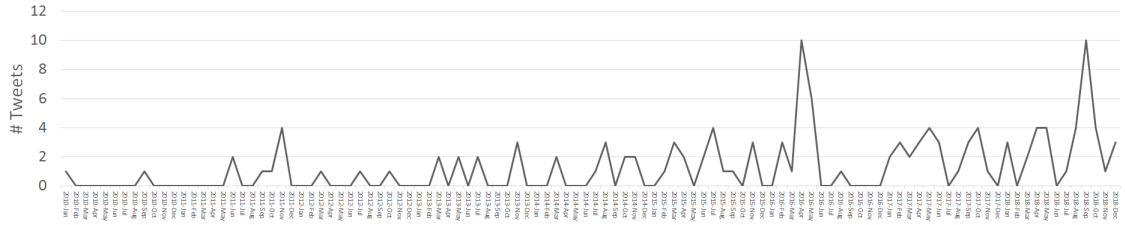

Figura 3.9: Distribuição temporal de tweets referentes a **Dermatite Atópica** entre 2010 e 2018, localizados em **Portugal**.

3.4 Redefinição de Objetivos

Através destas primeiras análises é claro que nem todos tweets são da autoria de doentes, nem todos se referem literalmente ao nome das doenças. Muitos se referem a celebridades e outros a publicidade a clínicas, eventos, notícias e programas de televisão. Ou seja, para um estudo do impacto das doenças na vida das pessoas, é necessário ir mais longe e explorar formas de separar os tweets que realmente são de interesse para este estudo (autoria do doente com doença referida no sentido literal) daqueles que, neste contexto, constituem ruído (publicidade, sentido figurado). Dada a quantidade de dados, isto deve ser feito de forma o mais automatizada possível, contemplando também a tarefa de reconhecimento emocional, já que os léxicos de texto não são resilientes às tantas palavras próprias da gíria característica das redes sociais. Torna-se necessário um sistema que contactando com parte das instâncias aqui reunidas, obtenha um poder de generalização para avaliar as restantes. Surge assim a necessidade de criar um conjunto de dados anotado para criar soluções de natureza supervisionada.

Capítulo 4

Classificação de tweets com base em Crowdsourcing

Neste capítulo, é tratada a projeção de um conjunto de tweets anotados, com recurso a uma plataforma on-line de *crowdsourcing*, na tentativa de encontrar uma solução de filtragem de tweets através de métodos de classificação supervisionada, para reduzir o conjunto total de cada doença ao subconjunto de interesse, i.e., aqueles onde o autor se pronuncia sobre a sua doença no sentido literal.

Dada a considerável limitação dos métodos não-supervisionados, como a análise de léxicos, em responder em todos os contextos, uma vez que foi decidido recorrer a anotação de tweets, foi pedido aos colaboradores que os catalogassem também segundo as coordenadas emocionais: valência e ativação.

Devido às limitações dos métodos não-supervisionados para identificar tweets de interesse em todos os contextos (p.ex, com base em palavras chave) e de identificar as principais emoções expressas (p.ex., com base em léxicos de emoção), optámos por adotar uma abordagem baseada em aprendizagem supervisionada para este fim. Este tipo de abordagem precisa de um conjunto inicial de dados, anotados de acordo com o objetivo. Como a criação desses dados por apenas uma pessoa seria uma tarefa morosa e, possivelmente, tendenciosa, recorremos a uma plataforma onde colaboradores voluntários anotaram tweets de acordo com um conjunto de parâmetros, que nos permitiriam, entre outros, perceber se os tweets eram de interesse para o estudo, bem como a emoção predominante.

Analisamos aqui os resultados da anotação, a distribuição pelas várias classes e a necessidade de adaptações. Segue-se o treino de classificadores, testados em vários tipos de *features*, e o recurso a validação cruzada para obter métricas de desempenho. Estas constituirão o critério para decidir a melhor combinação de *features* e algoritmo de classificação para cada problema abordado. Por fim, são obtidos resultados para duas modalidades distintas de filtragem de tweets, utilizando em comum as mesmas soluções de análise emocional. Considerando a falta de robustez nas soluções encontradas e na dificuldade em adquirir novas anotações via *crowdsourcing*, é procurada uma alternativa, abordada no capítulo seguinte.

4.1 Estrutura e resultados da anotação

Chegando à necessidade de obter um mecanismo para diminuir o ruído presente no conjunto de tweets, como é caso da publicidade ou uso metafórico na referência à doença, optou-se por uma solução baseada em classificação automática. Pela necessidade de anotar uma grande quantidade de tweets (várias centenas) e não viciar os dados com a opinião individual do autor, foi decidido recorrer a uma plataforma online de *crowdsourcing*, onde aos participantes é requerida a tarefa de anotação em troca de uma pequena quantia de dinheiro. A plataforma dispõe de métodos de controlo para avaliar o desempenho dos anotadores, como o intervalo de tempo entre anotações consecutivas, ou testes de atenção, como a anotação de um problema óbvio cuja resposta é absolutamente consensual ou limitada a um pequeno grupo de opções. Se o anotador incorre em falha pode mesmo ser excluído da tarefa.

A plataforma escolhida para anotação via *crowdsourcing* tem por nome ***Figure Eight***[1], e foi escolhida por ter sido utilizada anteriormente, pelo orientador, com algum sucesso.

Iniciando com alguns testes, foram enviados perto de uma centena de tweets que mencionam "psoríase", publicados durante a primeira metade de 2018. Não viriam a ser descartados, sendo que talvez tivesse sido mais apropriado selecionar aleatoriamente por todo o conjunto reunido dos vários anos. Ainda assim, confere aos classificadores um melhor discernimento sobre os tweets mais recentes, uma vez que as redes sociais são caracterizadas por uma gíria muito volátil, que se renova rapidamente.

Em conjunto com o orientador, foram definidas as perguntas que deveriam ser respondidas acerca de cada tweet. Essas perguntas seriam depois inseridas na plataforma, para que fossem respondidas pelos colaboradores. Mais propriamente, para cada tweet, pedíamos a resposta de três colaboradores diferentes às seguintes seis questões, colocadas após a apresentação do tweet, e com as opções que lhes seguem:

1. *Em que língua está escrita a mensagem?*

   Respostas possíveis: *Português, Outra*

2. *A mensagem será da autoria de uma pessoa individual ou de um grupo/organização?*

   Respostas possíveis: *Individual, Grupo/Organização, Não Sei*

3. *Na mensagem, a doença é referida num sentido:*

   Respostas possíveis: *Literal, Figurado*

4. *No caso de ser possível identificar, quem será o portador da doença referida?*

   Respostas possíveis: *Autor da mensagem, Outra pessoa, Não aplicável*

5. *O sentimento predominante da mensagem é:*
   Muito desagradável $\leftarrow$ *1, 2, 3, 4, 5* $\rightarrow$ *Muito agradável*

   Respostas possíveis: *1, 2, 3, 4, 5*

6. *O sentimento transmitido na mensagem é:*
   Nada intenso $\leftarrow$ *1, 2, 3, 4, 5* $\rightarrow$ *Muito intenso*

   Respostas possíveis: *1, 2, 3, 4, 5*

[1] `https://www.figure-eight.com/`

Da questão 1 pretendemos identificar que tweets enviados para a plataforma não estão escritos em língua portuguesa, para mais tarde serem excluídos do conjunto de treino para classificação, já que possivelmente, os tweets estarão incorretamente anotados, pois ao anotador não é exigido o conhecimento de outras línguas que não a portuguesa e, eventualmente, as features obtidas a partir desses tweets podem prejudicar os modelos. Ainda que os tweets tenham sido obtidos através de uma pesquisa que inclui uma opção de restrição para se obterem apenas tweets em língua portuguesa, o Twitter devolve algumas publicações em línguas estrangeiras.

Com a pergunta 2 pretendemos separar tweets publicados por apenas uma pessoa de outros publicados por contas de grupos, organizações (possivelmente publicidade) ou até notícias.

Principalmente no caso da urticária, é percetível o uso abundante do nome da doença para efeitos de metáfora, como em comentários políticos e outras críticas. Para discernir estes casos foi feita a pergunta 3.

Com a questão 4, pretendemos identificar o eventual portador da doença mencionada no tweet. Alguns utilizadores revelam que têm a doença, mas outras vezes referem-se a familiares, amigos ou conhecidos. A opção 'Não aplicável' está incluída para identificar os tweets onde nenhum doente é evidenciado, sendo que existem tweets cujo autor é individual e se refere à doença no sentido literal, mas não se refere a si ou a outra pessoa como doente.

Com a pergunta 5, o objetivo será identificar uma das dimensões que pode qualificar uma emoção: a valência. O mesmo para a pergunta 6, neste caso a ativação.

Antes de se submeterem tweets para a plataforma, deve-se ter o cuidado de evitar repetições, já que o mesmo tweet, se mencionar mais do que uma das doenças, estará presente em vários conjuntos. Pretende-se por isso evitar a redundância, já que o mesmo tweet é anotado por pelo menos três pessoas. Mesmo dentro do conjunto da mesma doença, existem tweets que, embora não se tratem de partilhas, excluídas antes da obtenção através de um modificador na pesquisa (*-filter:retweets*), são publicados a partir de páginas externas com o mesmo texto predefinido. Segue-se um exemplo, onde a única diferença reside no sufixo do endereço URL:

```
| VITILIGO: MICHAEL JACKSON NÃO FOI O ÚNICO | ‘‘Rara doença de pele
transformou Darcel Vlugt de uma criança...  https://t.co/Nq8ocjOR4a
```

```
| VITILIGO: MICHAEL JACKSON NÃO FOI O ÚNICO | ‘‘Rara doença de pele
transformou Darcel Vlugt de uma criança...  https://t.co/y2ZMQZhMxs
```

A identificação de casos como o anterior foi possível após a substituição de todos os URLs para um identificador comum `URLREF`. Isto acabou também por aumentar o poder de discernimento dos modelos de classificação, ao contabilizar a presença de endereços URL no tweet. Assim o prova uma análise de relevância de features baseada no teste estatístico do χ^2, que a atesta como a feature mais relevante em todos os problemas de classificação, exceto o sentido (literal, figurado). De facto, por inspeção manual, a presença de endereços URLs é mais frequente em tweets de contas não pessoais, estando presentes em 10% dos tweets anotados como tendo o autor como doente, em 53% dos tweets que apontam o doente como sendo outra pessoa, e 77% dos tweets identificados como não tendo menção a doentes.

Nas tabelas 4.1 e 4.2 são apresentados os resultados das anotações: a distribuição pelas

várias classes, respetivamente em frequência absoluta e percentagem. Estão contabilizados 478 tweets anotados (100 de psoríase, 89 de urticária e 289 de dermatite atópica), já excluindo os 11 identificados com língua estrangeira. A grande quantidade de tweets que mencionam "dermatite atópica" em relação às restantes, deve-se a uma potencial colaboração com um projeto relacionado a esta doença, que acabou por não se concretizar.

#	**Autor**			**Sentido**		**Doente**			**Binário**	
	Individual	**Grupo**	**NA**	**Literal**	**Figurado**	**Autor**	**Outra**	**NA**	**ILA**	**Não-ILA**
Psoríase	71	10	19	96	4	52	7	41	50	50
Urticária	77	5	7	45	44	55	7	27	29	60
D.Atópica	209	62	18	285	4	143	48	98	140	149
TOTAL	357	77	44	426	52	250	62	166	219	259

Tabela 4.1: Distribuição dos dados anotados via **crowdsourcing** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores.

%	**Autor**			**Sentido**		**Doente**			**Binário**	
	Individual	**Grupo**	**NA**	**Literal**	**Figurado**	**Autor**	**Outra**	**NA**	**ILA**	**Não-ILA**
Psoríase	71.0	10.0	19.0	96.0	4.0	52.0	7.0	41.0	50.0	50.0
Urticária	86.5	5.6	7.9	50.6	49.4	61.8	7.9	30.3	32.6	67.4
D.Atópica	72.3	21.5	6.2	98.6	1.4	49.5	16.6	33.9	48.4	51.6
TOTAL	74.7	16.1	9.2	89.1	10.9	52.3	13.0	34.7	45.8	54.2

Tabela 4.2: Distribuição em **percentagem** dos dados anotados via **crowdsourcing** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores.

Os agrupamentos de colunas respetivos a autor, sentido e doente, provêm diretamente das anotações via *crowdsourcing*. O último agrupamento ('Binário'), foi gerado pelo autor com base nos dados anotados. O acrónimo 'ILA' é aqui utilizado para identificar os tweets cujo autor é **I**ndividual, o sentido da referência à doença é **L**iteral e o doente é o **A**utor.

Da tabela 4.2, podemos concluir que a vasta maioria dos tweets anotados tem autor individual. Quanto ao sentido, metade dos tweets que referem 'urticária', o fazem no sentido figurado. Na totalidade dos tweets ('TOTAL'), 11% foram anotados assim. Dermatite atópica tem o dobro da proporção dos tweets que mencionam outra pessoa doente, relativamente a psoríase e urticária. Metade do conjuntos anotados relativos a psoríase e dermatite atópica são tweets ILA. Já no caso de urticária a proporção é de apenas um terço.
A binarização dos dados permite, neste caso, a vantagem de se obter um conjunto de treino consideravelmente balanceado. No entanto, nesta fase ainda é desconhecido se fornecerá informação suficiente ao algoritmo de classificação para o tornar resiliente ao sentido figurado presente em tantos tweets.

À semelhança das tabelas abordadas anteriormente, as tabelas 4.3 e 4.4 contêm dois agrupamentos de classes ('Valência Simples' e 'Ativação Simples') gerados posteriormente a partir dos dados anotados de 'Valência' e 'Ativação'.

#	**Valência**					**Ativação**					**Val. Simples**			**Ati. Simples**		
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**-**	**0**	**+**	**-**	**0**	**+**
Psoríase	15	29	34	7	3	2	14	32	23	15	52	34	14	22	32	46
Urticária	10	26	50	2	1	3	6	59	16	5	36	50	3	9	59	21
D.Atópica	65	84	114	16	4	3	6	123	101	49	155	114	20	11	123	155
TOTAL	90	139	198	25	8	8	26	214	140	69	243	198	37	42	214	222

Tabela 4.3: Distribuição dos dados anotados via **crowdsourcing** pelos vários níveis de valência e ativação (1 a 5), e nas classes geradas posteriormente (simplificação para três labels possíveis: Negativo (-), Origem (0), Positivo(+)).

Valência e ativação são tratados tipicamente como problemas de regressão. No entanto,

%	Valência					Ativação					Val. Simples			Ati. Simples		
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**-**	**0**	**+**	**-**	**0**	**+**
Psoríase	15.0	29.0	34.0	7.0	3.0	2.0	14.0	32.0	23.0	15.0	52.0	34.0	14.0	22.0	32.0	46.0
Urticária	11.2	29.2	56.2	2.2	1.1	3.4	6.7	66.3	18.0	5.6	40.4	56.2	3.4	10.1	66.3	23.6
D.Atópica	22.5	29.1	39.4	5.5	1.4	1.0	2.1	42.6	34.9	17.0	53.6	39.4	6.9	3.8	42.6	53.6
TOTAL	18.8	29.1	41.4	5.2	1.7	1.7	5.4	44.8	29.3	14.4	50.8	41.4	7.7	8.8	44.8	46.4

Tabela 4.4: Distribuição em **percentagem** dos dados anotados via **crowdsourcing** pelos vários níveis de **valência** e **ativação** (1 a 5), e nas classes geradas posteriormente (simplificação para três labels possíveis: Negativo (-), Origem (0), Positivo (+)).

devido à complexidade do problema e ao número de dados disponíveis, os resultados de testes com modelos de regressão não foram satisfatórios. Apesar de tudo, para a análise a realizar, não seria necessário uma classificação tão granular para a emoção, por isso foi feita uma modificação que torna o problema mais adequado à classificação, com melhores resultados, e também perceção conceptual: com cinco níveis para valência e ativação, dividimos o referencial emocional em 25 regiões (5×5), depois simplificadas para 9 regiões. Cada região não representa uma emoção, mas várias. Por exemplo, o quadrante de valência negativa e alta ativação (V-A+) é caracterizado por emoções como raiva, ansiedade e frustração. Está a ser aplicado o modelo de Russell (1980), mas de uma forma mais simples, considerando-se apenas quadrantes, zonas intermédias e neutras, de forma mais semelhante ao que é feito por Malheiro et al. (2016).

A simplificação seguiu o seguinte critério, de valência/ativação (esquerda) para valência simples/ativação simples (direita):

- $[1, 3[\rightarrow -$
- $[3] \rightarrow 0$
- $]3, 5] \rightarrow +$

Note-se que as instruções de transformação fazem o uso de intervalos, e não de valores discretos. A razão prende-se com o facto de alguns tweets não terem, para valência ou ativação, um número inteiro. Cada tweet foi anotado por várias pessoas e, naturalmente, ocorrem anotações distintas. O primeiro critério a utilizar para definir o valor final para cada uma das coordenadas foi escolher a moda do conjunto de anotações do tweet. Isto privilegiaria o consenso entre os anotadores e a resiliência contra uma eventual distração de um outro, que logo influenciaria o resultado final se o critério primeiro consistisse no cálculo da média. Surgindo a possibilidade de ocorrer um empate entre valores, não sendo possível definir a moda, foi calculada a mediana dos valores empatados. Só no caso de persistirem dois valores distintos como mediana (se o conjunto for par), foi calculada a sua média, podendo originar números não-inteiros. A ocorrência destes casos não está contemplada nos dois primeiros agrupamentos (Valência, Ativação) nas tabelas 4.3 e 4.4, mas está presente nos agrupamentos da simplificação a três valores ('Val. Simples','Ati. Simples').

Podemos observar que o conjunto de treino é fraco em exemplares com valência positiva (7.7%) e baixa ativação (8.8%), o que eventualmente condicionará os resultados.

As tabelas 4.5 e 4.6, resultam da combinação das coordenadas de valência e ativação (simples). Cada coluna corresponde a um quadrante, fronteira ou à origem do referencial emocional adotado. A proporção de tweets de valência e ativação neutras (V0A0) e valência negativa e alta ativação (V-A+) são bastante consideráveis.

#	Referêncial Emocional								
	V+A-	**V+A0**	**V+A+**	**V0A-**	**V0A0**	**V0A+**	**V-A+**	**V-A0**	**V-A-**
Psoríase	2	3	9	6	24	4	33	5	14
Urticária	0	3	0	4	40	6	15	16	5
D.Atópica	2	12	6	7	94	13	136	17	2
TOTAL	4	18	15	17	158	23	184	38	21

Tabela 4.5: Distribuição dos dados anotados via **crowdsourcing** pelos vários **quadrantes** e fronteiras do referêncial emocional adoptado.

%	Referêncial Emocional								
	V+A-	**V+A0**	**V+A+**	**V0A-**	**V0A0**	**V0A+**	**V-A+**	**V-A0**	**V-A-**
Psoríase	2.0	3.0	9.0	6.0	24.0	4.0	33.0	5.0	14.0
Urticária	0.0	3.4	0.0	4.5	44.9	6.7	16.9	18.0	5.6
D.Atópica	0.7	4.2	2.1	2.4	32.5	4.5	47.1	5.9	0.7
TOTAL	0.8	3.8	3.1	3.6	33.1	4.8	38.5	7.9	4.4

Tabela 4.6: Distribuição em **percentagem** dos dados anotados via **crowdsourcing** pelos vários **quadrantes** e fronteiras do referencial emocional adotado.

4.2 Obtenção dos classificadores

Como já descrito no capítulo 3, um dos passos do pré-processamento englobou a substituição dos termos relacionados à doença por `DISEASEREF`. Esta é uma prática comum em tarefas de classificação. Por exemplo, com o objetivo de criar um sistema que detete menções ao estado de saúde pessoal em tweets, Yin et al. (2015) conseguiu, a partir de um conjunto de treino de 2000 tweets mencionando uma de quatro doenças, obter um classificador com precisão de 0.77, quando testado num conjunto de dados de 34 doenças distintas. Para tal, processou o texto de forma a melhorar a generalização do classificador.

Os endereços URLs não eram necessários conservar, sendo que podem ser acedidos a qualquer momento, mas não se deduz que os algoritmos de classificação beneficiem com a estrutura do endereço. A sua presença, no entanto, é muito significativa e oferece muito poder de discernimento.

Depois de várias experiências de adaptação à biblioteca **Scikit-Learn**, foi decidido selecionar um conjunto de algoritmos suficientemente diferentes, começando pelos métodos mais tradicionais e perceber o que é possível fazer com eles. Segue-se o conjunto selecionado (informação no capítulo 2):

- **MultinomialNB**: Multinomial Naive Bayes
- **LinearSVC**: Uma implementação de Support Vector Machine para Classificação.
- **DesisionTree**: Árvore de Decisão, critério: Gini
- **NearestCentroid**: Classificador Nearest Centroid, distância: euclidiana

Procedeu-se também à averiguação da variedade de features e cuidados a incluir:

- **Tipo de termos**: Tokens (originais) vs Lemas
- **N-gramas**: N=1 vs N=2 vs N=1,2 vs N=1,2,3
- **Vectorização**: Count (Frequência Absoluta) vs TF (Frequência Relativa) vs TF-IDF (Frequência Relativa multiplicada pelo Inverso da Frequência nos Documentos)

- **Stop-words**: Remover ou não
- **Acentos**: Remover ou não

Todas as combinações de classificadores e tipos de features foram avaliadas para cada problema de classificação, recorrendo a *10-fold cross-validation*. Os melhores resultados, usando como critério a medida F1-pesada, estão listados na tabela 4.7. Foram usados os paramentros padrão em todos os classificadores.

Como dito anteriormente, também foi averiguada a possibilidade de proceder à filtragem de tweets com um classificador binário (ILA vs não-ILA).

Com o objetivo de conhecer qual o tipo de features a utilizar, foi realizado um conjunto de experiências a partir de *10-fold cross-validation*, obtendo-se a média da medida F1 como critério de escolha da solução a implementar em cada problema, versão *weighted* para considerar o desbalanceamento do conjunto de dados.

O problema da valência e ativação pode ser abordado segundo a lógica de regressão. As abordagens retratadas seguem a lógica de classificação: os valores foram tratados como rótulos e não como números. Seguindo esta lógica, uma vez que os resultados obtidos são muito insatisfatórios, foi utilizada a simplificação para 3 *labels*, que anteriormente foi nomeada como *Valência Simples* e *Ativação Simples*. Doravante serão referidas apenas por *Valência* e *Ativação*. Os problemas de 5 níveis (1 a 5) serão referidos com *Valência-5* e *Ativação-5*. O conjunto dos melhores classificadores para cada problema encontra-se listado na tabela 4.7.

Classificação	**Modelo**	**Lemas**	**N-gramas**	**Vectorização**	**Remoção Stop-words**	**Remoção Acentos**	**Número Features**
Autor	LinearSVC	Não	1	TF	Não	Sim	697
Sentido	LinearSVC	Sim	1-2	Count	Não	Não	1481
Doente	LinearSVC	Não	1-2	TF	Não	Sim	1498
Class. Binário	LinearSVC	Não	1	TF-IDF	Não	Não	765
Valência	NearestCentroid	Sim	1-2	TF-IDF	Não	Não	1481
Ativação	NearestCentroid	Sim	1-2	TF-IDF	Não	Sim	1488
Valência-5	NearestCentroid	Não	1-2	TF-IDF	Não	Não	1500
Ativação-5	NearestCentroid	Sim	1-2	TF-IDF	Não	Não	1481

Tabela 4.7: Melhor combinação de características para cada problema de classificação. Conjunto de treino: *crowdsourcing*. *Class. Binário* corresponde ao classificador binário que separa os tweets em cujo autor é o doente e fala no sentido literal, dos restantes.

Todas as combinações de caraterísticas e algoritmos de classificação (385 para cada problema), foram avaliadas através de *10-fold cross-validation*, de onde foram obtidas métricas de performance. Produziram-se listas dos resultados para cada problema de classificação. Uma vez ordenadas segundo F1 (pesada), mostram que a maioria das combinações dentro do mesmo problema, atinge um empate técnico em todas as métricas quando considerado o desvio-padrão. Da mesma forma, não é evidente uma prevalência de um algoritmo de classificação ou de certo tipo de *features* nos primeiros lugares, com excepção da preservação de stop-words, comum em todos os problemas, e a não utilização de *Lemas* nos classificadores de *Doente* e *Classificador Binário*(*ILA vs não-ILA*), comum entre pelos menos os primeiros 15 lugares.

Contudo, quando o melhor resultado de cada problema é listado numa tabela (4.7), podemos concluir que todos os problemas correspondentes à tarefa de filtragem de tweets (os quatro primeiros listados), têm *LinearSVC* como algoritmo de classificação, uma implementação de SVM.

Os classificadores selecionados para a tarefa de reconhecimento emocional (incluindo os já rejeitados, Valência-5 e Ativação-5) também têm em comum o mesmo algoritmo de classificação, a implementação *NearestCentroid*. Para além disso, têm também em comum a utilização de unigramas e bigramas, bem como a vectorização *TF-IDF*. Destaca-se ainda, que ambos os classificadores de *Valência* e *Ativação* a serem usados, fazem uso de *Lemas*. Segue-se de seguida a lista das métricas de performance para cada problema (tabela 4.8).

Classificação	**Accuracy**	**Precision**	**Recall**	**F1**	**F1 (pesado)**
Autor	0.949 ± 0.035	0.918 ± 0.065	0.916 ± 0.064	0.912 ± 0.054	0.949 ± 0.035
Sentido	0.921 ± 0.047	0.837 ± 0.103	0.765 ± 0.124	0.776 ± 0.087	0.913 ± 0.056
Doente	0.828 ± 0.053	0.835 ± 0.067	0.752 ± 0.092	0.767 ± 0.087	0.820 ± 0.060
Class. Binário	0.835 ± 0.037	0.838 ± 0.036	0.838 ± 0.034	0.832 ± 0.037	0.835 ± 0.037
Valência	0.734 ± 0.069	0.689 ± 0.113	0.658 ± 0.127	0.647 ± 0.104	0.725 ± 0.073
Ativação	0.688 ± 0.054	0.525 ± 0.067	0.546 ± 0.105	0.520 ± 0.066	0.674 ± 0.064
Valência-5	0.523 ± 0.054	0.294 ± 0.058	0.296 ± 0.053	0.287 ± 0.048	0.519 ± 0.059
Ativação-5	0.546 ± 0.065	0.292 ± 0.042	0.284 ± 0.044	0.277 ± 0.036	0.548 ± 0.059

Tabela 4.8: Performance dos classificadores selecionados a partir de *10-fold cross-validation*. Precision-macro, Recall-macro, F1-macro e F1-weighted. Conjunto de treino: **crowdsourcing**.

Existem duas possíveis abordagens para filtrar os tweets de interesse, ou seja, identificar os tweets ILA, aqueles cujo autor é *Individual*, o sentido da referência à doença é *Literal* e onde efetivamente é mencionado um doente que é o *Autor*. A primeira opção é aplicar, a cada conjunto de tweets por doença, os três classificadores (Autor, Sentido e Doente) e no final extrair aqueles que incluem as três labels pretendidas.

A segunda opção é aplicar o classificador binário (*ILA vs-não-ILA*), que acaba por utilizar a informação disponível em cada uma das classes anteriores, embora não seja conhecido se o poder de discernimento se reduz numa tarefa específica, como a questão do uso figurado.

Dados os resultados na tabela 4.8, especificamente a coluna F1 (não-pesado), o valor correspondente ao *classificador binário* é superior ao de *Sentido* e *Doente*. Uma vez que a primeira abordagem é cruzar os resultados do classificador de *Autor* com os classificadores de *Sentido* e *Doente*, a performance final fica condicionada pelo pior desempenho entre os três classificadores. Neste caso trata-se de *Doente*, cuja medida F1-pesada (0.820±0.060) é semelhante à do classificador binário (0.835±0.037).
As restantes métricas parecem beneficiar o classificador binário, com exceção da classificação de *Autor*. Ainda assim, será obtida a apuração de tweets de interesse *ILA vs não-ILA*, para ambas as modalidades, e talvez a análise de vocabulário sugira qual a mais apropriada. A apuração de emoções será a mesma para ambos os casos, sendo aplicados os mesmos classificadores de *Valência* e *Ativação*.

O conjunto de treino vai ser reduzido para a classificação de *Autor*, excluindo os tweets com label '*NA*', de forma a classificar cada tweet como tendo um autor *individual* ou um grupo responsável pela sua publicação. O mesmo não será feito no problema da identificação do *Doente*, já que '*NA*' é uma opção válida, pois não haver nenhum doente mencionado no tweet é uma opção real e bastante frequente. As tabelas 4.9 e 4.10 apresentam os resultados da aplicação das duas modalidades nos conjuntos das doenças, incluindo a correspondência entre elas, ou seja, o número/percentagem de tweets com classificações idênticas.

Da tabela 4.11, podemos observar que a menor correspondência (número/percentagem) de previsões iguais entre ASD e Classificador Binário) acontece para urticária (62.6%). Tal

#	Autor		Sentido		Doente			Class. Binário	
	Indiv.	**Grupo**	**Literal**	**Figurado**	**Autor**	**Outra**	**NA**	**ILA**	**não-ILA**
Psoríase	14.257	8.094	22.165	186	6.039	254	16.058	4.769	17.582
Urticária	19.513	2.436	20.031	1.918	16.531	187	5.231	8.059	13.890
D. Atópica	3.255	1.159	4.368	46	1.985	112	2.317	1.613	2.801
Vitiligo	27.326	28.398	55.174	550	15.513	1.138	39.073	9.780	45.944
Diabetes	378.616	194.386	557.938	15.064	187.143	8.158	377.701	120.152	452.850
Asma	381.189	36.801	409.564	8.426	31.5312	8.511	94.167	232.003	185.987

Tabela 4.9: Distribuição de tweets que mencionam cada doença entre várias classes.

%	Autor		Sentido		Doente			Class. Binário	
	Indiv.	**Grupo**	**Literal**	**Figurado**	**Autor**	**Outra**	**NA**	**ILA**	**não-ILA**
Psoríase	63.8	36.2	99.2	0.8	27.0	1.1	71.8	21.3	78.7
Urticária	88.9	11.1	91.3	8.7	75.3	0.9	23.8	36.7	63.3
D. Atópica	73.7	26.3	99.0	1.0	45.0	2.5	52.5	36.5	63.5
Vitiligo	49.0	51.0	99.0	1.0	27.8	2.0	70.1	17.6	82.4
Diabetes	66.1	33.9	97.4	2.6	32.7	1.4	65.9	21.0	79.0
Asma	91.2	8.8	98.0	2.0	75.4	2.0	22.5	55.5	44.5

Tabela 4.10: Distribuição percentual de tweets que mencionam cada doença entre várias classes.

Doença	Classificadores ASD				Class. Binário				Correspondência	
	ILA	**%**	**Não-ILA**	**%**	**ILA**	**%**	**Não-ILA**	**%**	**#**	**%**
Psoríase	5898	26.4	16453	73.6	4769	21.3	17582	78.7	20282	90.7
Urticária	14726	67.1	7223	32.9	8059	36.7	13890	63.3	13736	62.6
D. Atópica	1940	44.0	2474	56.0	1613	36.5	2801	63.5	3845	87.1
Vitiligo	15044	27.0	40680	73.0	9780	17.6	45944	82.4	48008	86.2
Diabetes	173946	30.4	399056	69.6	120152	21.0	452850	79.0	486492	84.9
Asma	307195	73.5	110795	26.5	232003	55.5	185987	44.5	318170	76.1

Tabela 4.11: Comparação de resultados de filtragem de tweets ILA, por duas modalidades: Classificadores ASD (cruzamento de resultados dos classificadores individuais de Autor, Sentido e Doente) e Classificador Binário.

se deve prender com a questão do uso figurado. Por um lado, o classificador de *Sentido* é específico para esta tarefa, mas por outro, o classificador binário beneficia por também incluir a informação do tipo de autor e doente numa única *label*.

Nos restantes tópicos deste capítulo apresentam-se os resultados provenientes das duas modalidades de filtragem de tweets ILA, sob a forma de análise de tópicos, acompanhada do exercício especulativo para compreender os assuntos representados. Segue-se depois os resultados da aplicação dos classificadores de emoção, e a distribuição ao longo do tempo mais pormenorizada.

4.3 Resultados da filtragem com classificador binário

Nesta secção, apresentam-se os resultados relativos a tweets classificados como ILA pelo *Classificador Binário (ILA vs não-ILA)*. Primeiramente, pretende-se compreender os principais assuntos presentes na reunião dos subconjuntos ILA das quatro dermatoses (psoríase, urticária, dermatite atópica e vitiligo). Nesta análise não foram consideradas as doenças de diabetes e asma, uma vez que tão grande número de dados ($\approx$ 350 mil tweets) não daria espaço às doenças que constituem o foco principal deste estudo ($\approx$ 24 mil tweets). Depois são apresentados os três conjuntos de gráficos circulares. O primeiro representa as diferentes proporções das emoções em cada subconjunto de cada doença. Os dois seguintes correspondem a uma análise mais fina do primeiro. Por fim, os tweets são agrupados por tipo de emoção e distribuídos ao longo do tempo.

4.3.1 Análise de Tópicos

Nesta subsecção são apresentados 20 tópicos obtidos da reunião dos subconjuntos ILA das doenças psoríase, urticária, dermatite atópica e vitiligo, a partir da aplicação da técnica de LDA. Ao extrair os assuntos (tópicos) que compõem o documento, neste caso o conjunto de tweets ILA, pretende-se, após um exercício algo especulativo que inclui também alguma consulta direta ao conjunto de dados, perceber o significado do agrupamento de palavras que é apresentado por tópico e, comparar de forma global com os resultados obtidos no capítulo 3, onde se utilizaram os conjuntos originais. O objetivo é verificar se os novos tópicos obtidos sugerem uma evolução em relação aos tópicos originais, para perceber a eficácia do método de filtragem em resolver os problemas de ruído e contactar com as expressões dos doentes.

#	Tópico
0	mim sai urticaria meu rs jogo 😞 daqui celular olhar total pescoço outra aconteceu loca hahaha errado viada nós quê
1	me urticaria vitiligo não tu tengo ligo deixa outro vale dia prazer eu ligar mandar livre você lo mil URLREF
2	psoriase minha URLREF estresse sobre dermatite_atopica mais pele google vida voltando 😒 atacando mulheres não dr auto psoríase doença rinite
3	vitiligo eu tenho parece to meu descascando acho tô rosto parecendo cara minha estou sol tão agora pele manchas branca
4	eu vitiligo minha não tenho psoriase me meu sou urticaria nao ela mais muito mãe tava agora to fico bem
5	urticaria eu alergia minha tô estou nervosa tenho crise agora não calor 3 frio 2 alérgica meu 1 ataca falta
6	voltou psoriase vcs sabem voltar minha não atacar dado ideia fazia 😑 virar dentro bonita vocês muitas erro solução piorou
7	😂 merece ninguém mata 😫 🤧 29 chamado hahahah clube neh espera luis tt outubro horrível nois razão explicação curte
8	não vitiligo URLREF psoriase cura doença pele me dermatite_atopica tenho minha eu mais tratamento posso preconceito pessoas manchas seja contagiosa
9	pomada psoriase 😩 chama viver ponto <3 negócio daquele dermatologista tao aumentando alto noção sexta beijo nao bizarro forte vcs
10	anos dermatite_atopica tenho há aos 10 😣 hidratante coceira urticaria 15 remédio chegou médica sobre dias uso médico descobri menina
11	psoriase meu anos vitiligo URLREF 5 desde deus 😢 cu oi dermatite_atopica tomar meninas acabar 6 7 minha idade tenho
12	urticaria 😇 leio 😌 fala pensa 😪 Twitter detesto etc matando pros esquerda juro assistir amor 😢 en ex teria
13	você aquele semana fdp noite fim força :'(dormi volta chegando 😡 psoriase passada manhã meia comigo espinhas estudar 17
14	:(psoriase efeito kim vitiligo kardashian chato tenho entao comum infelizmente aquagênica af alergica dedo ❤️ mentira sou queimadura
15	vitiligo michael jackson era ele não branco URLREF mais negro doença pele mj eu usava meu preto peruca sabia filho
16	😭 😢 ótimo música rápido time adivinha própria inchaço leite casos man decidi ignorância ajudando belo língua ataques pego completamente
17	psoriase não eu minha urticaria dermatite_atopica mais meu dia bom atacada corpo aguento porra hoje tô merda to pior nao
18	mancha vitiligo odeio limão parece disse eu limao perguntaram 8 mais todas forças sofria piadas crescer cacete esperando banda rolê
19	urticaria me la puta provoca chega produce 😍 causa tu dando con maldita D= el dio facebook vejo merda dormir

Tabela 4.12: 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelo **classificador binário** treinado com os dados de **crowdsourcing**.

O tópico 0 é de difícil compreensão. Destaca-se a ocorrência de pronomes na primeira pessoa do singular ('mim', 'meu') e plural ('nós'). Contém um emoji que significa decepção, bem como 'rs' (risos) e 'hahaha' (gargalhada).

No tópico 1 voltam a estar presentes pronomes ('me', 'tu', 'eu', 'você'). Demonstra ter em conta tweets em língua espanhola ('tengo', 'lo').

Os tópicos 0 e 1 mencionam features que são importantes para identificar tweets ILA,

e por isso é normal que sejam frequentes, principalmente depois da filtragem.

O tópico 2 menciona duas dermatoses (psoríase, dermatite atópica) e rinite. A dermatite atópica pode provocar uma rinite alérgica [2]. O tópico sugere a presença descrições do estado de saúde pessoal pela presença do pronome 'minha', compatível com 'psoríase', 'dermatite atópica', 'doença', 'rinite', 'pele' ou 'vida'. O tópico parece sugerir a volta ('voltando') dos sintomas da doença, invocando stress ('estresse'). Aparece ainda um emoji associado a uma emoção negativa, com o nome de 'unamused face' (sem graça).

O tópico 3 parece tratar da comparação das queimaduras solares ('sol', 'pele', 'descascando') com as 'manchas' brancas ('branca'), características de 'vitiligo'.

No tópico 4 são mencionadas três dermatoses (psoríase, vitiligo e urticária). Os pronomes presentes ('eu', 'minha', 'me', 'meu'), bem como as formas verbais ('tenho', 'sou', 'to') sugerem referências ao próprio autor, embora 'ela' e 'mãe' sugiram a referência a outra pessoa. O tópico 6 da tabela 3.3 também inclui 'ela', 'mãe' e 'boa'.

O tópico 5 expressa o estado nervoso ('nervosa') da autora, que poderá ter desencadeado uma 'crise' 'alérgica' de urticária. É mencionado 'calor' e 'frio', tratando-se talvez de urticária colinérgica ou de frio, como mencionado no tópico 18 da tabela 3.3.

O tópico 6 parece tratar do agravamento ('voltou', 'piorou') dos sintomas de psoríase. O emoji presente é denominado 'expressionless face', embora contenha uma conotação negativa, ou de algo pouco agradável acerca do qual não se pode fazer muito. O seu uso talvez esteja associado às consequências da doença na beleza ('bonita').

O tópico 7 contém três emojis, dois deles com óbvia conotação negativa, juntamente com a palavra 'horrível', em oposição à gargalhada ('hahahah') e à palavra 'curte' (gosto, like). Contém ainda a referência ao Dia Mundial da Psoríase (29 de Outubro), mas não menciona a doença, podendo tratar-se de um falso positivo.

O tópico 8 faz menção ao sofrimento causado pelo estigma social associado às dermatoses, pelo medo que sejam contagiosas ('contagiosa'), o que não passa de um 'preconceito'.

No tópico 9 está presente o termo 'pomada', que se refere provavelmente a um tratamento contra a psoríase. O tópico é bastante heterogéneo.

O tópico 10 faz menção a urticária e dermatite atópica, a sintomas ('coceira') e tratamento ('hidratante', 'remédio', 'médico', 'médica').

O tópico 11 é bastante heterogéneo. Inclui o recurso a emoji de emoção negativa e calão ('cu').

O tópico 12 parece tratar do recurso a urticária no sentido figurado ('pensa', 'detesto') no âmbito da discussão política ('leio', 'fala', 'pensa', 'esquerda') ou de um desgosto amoroso ('amor', 'ex'). O recurso a urticária no sentido figurado também se faz presente no tópico 14 da tabela 3.3, tratando-se de mais falsos positivos (ou seja, não-ILA). Estão presentes quatro emojis, três dos quais têm conotação negativa.

O tópico 13 sugere a lamentação de sofrer psoríase, acne ('espinhas') ou algum distúrbio durante o sono ('dormi'). Um emoji e um emoticon, ambos de conotação negativa, estão presentes.

O tópico 14 faz referência a Kim Kardashian, celebridade portadora de psoríase. Uma referência a urticária de alergia à água ('aquagénica') está provavelmente presente, já que, após pesquisa, não foi encontrada uma variante aquagénica da psoríase.

[2] `https://www.pfizer.com.br/noticias/Diferen%C3%A7a-entre-psoriase-e-dermatite-atopica`

O tópico 15 retrata das polémicas associadas a Michael Jackson, de forma semelhante ao tópico 17 da tabela 3.3.

Não foram encontradas pistas que justifiquem a diversidade de termos no tópico 16. Existem casos de urticária desencadeada por alergia ao leite.

O tópico 17 expressa, de forma óbvia, a lamentação do autor numa dermatose que atacou ('atacada') o seu 'corpo', não aguentando a situação ('não', 'aguento'), expressando o seu descontentamento na forma de calão ('porra', 'merda').

O tópico 18 é bastante heterogéneo, fazendo referência a 'limão', talvez como tratamento natural contra vitiligo, à semelhança do tópico 17 na tabela 3.3.

À semelhança do tópico 14 da tabela 3.3, o tópico 19 é influenciado por tweets em língua espanhola ('la', 'produce', 'con', 'el', 'dio'). A presença de 'provoca' sugere o uso figurado de urticária. O tópico contém calão ('puta', 'merda'), um emoji positivo e um emoticon negativo.

Apreciação

Contam-se **5** tópicos com a presença do termo '`URLREF`' na tabela 4.12, menos quatro do que os listados na tabela 3.3. Isto sugere um menor número de ligações para páginas externas que normalmente não estão presentes em tweets que mencionam o estado de saúde do autor. Existir uma redução vai ao encontro do esperado. Da mesma maneira, é de esperar um maior recurso a emojis e emoticons quando se transmite o estado de saúde pessoal, não tão habituais quando se partilham artigos meramente informativos. Na tabela 4.12, contam-se 18 emojis (14 negativos) e 4 emoticons (3 negativos). Pode concluir-se assim, que a filtragem teve alguma eficácia. Pela análise especulativa, admitem-se **8** tópicos compatíveis com tweets ILA: 2, 4, 5, 6, 8, 10, 11 e 17. Por outro lado, é perceptível que a filtragem não terá sido completamente eficaz, já que alguns dos tópicos parecem continuar associados a tweets não desejados neste estudo. Isto mostra que a filtragem será, afinal, mais difícil do que sugerem os resultados da validação cruzada.

4.3.2 Emoções

A figura 4.1 apresenta os gráficos da distribuição emocional no conjunto e subconjuntos de cada doença. Cada cor corresponde ao agrupamento de emoções segundo o referencial emocional adotado. Observa-se a erradicação de emoções positivas (azul, azul-esverdeado, verde) nos subconjuntos ILA, podendo a sua presença ser desprezada. Apenas é percetível uma pequena linha a azul (V+A-) no caso de vitiligo. A maior semelhança, considerando todos os gráficos por doença, está entre vitiligo e diabetes. Os gráficos ILA encontram muito maior semelhança entre si do que os restantes. A diferença dos tons de cinzento não é relevante para a análise já que se tratam de regiões com valência neutra, o que não faz tanto sentido em falar de diferentes níveis de ativação.

Segue-se a ordenação decrescente das percentagens dos grupos emocionais de valência negativa dos subconjuntos ILA:

V-A+ (vermelho): dermatite atópica (65.8%), psoríase (60.5%), asma (60.2%), urticária (56.4%), diabetes (53.5%), vitiligo (53.2%).

V-A0 (laranja): asma (21%), vitiligo (20.2%), diabetes (18.6%), psoríase (16.5%), urticária (16.4%), dermatite atópica (16.1%).

V-A- (amarelo): vitiligo (7.4%), psoríase (6.7%), diabetes (5.7%), asma (5.3%),

dermatite atópica (5.2%), urticária (4.4%).

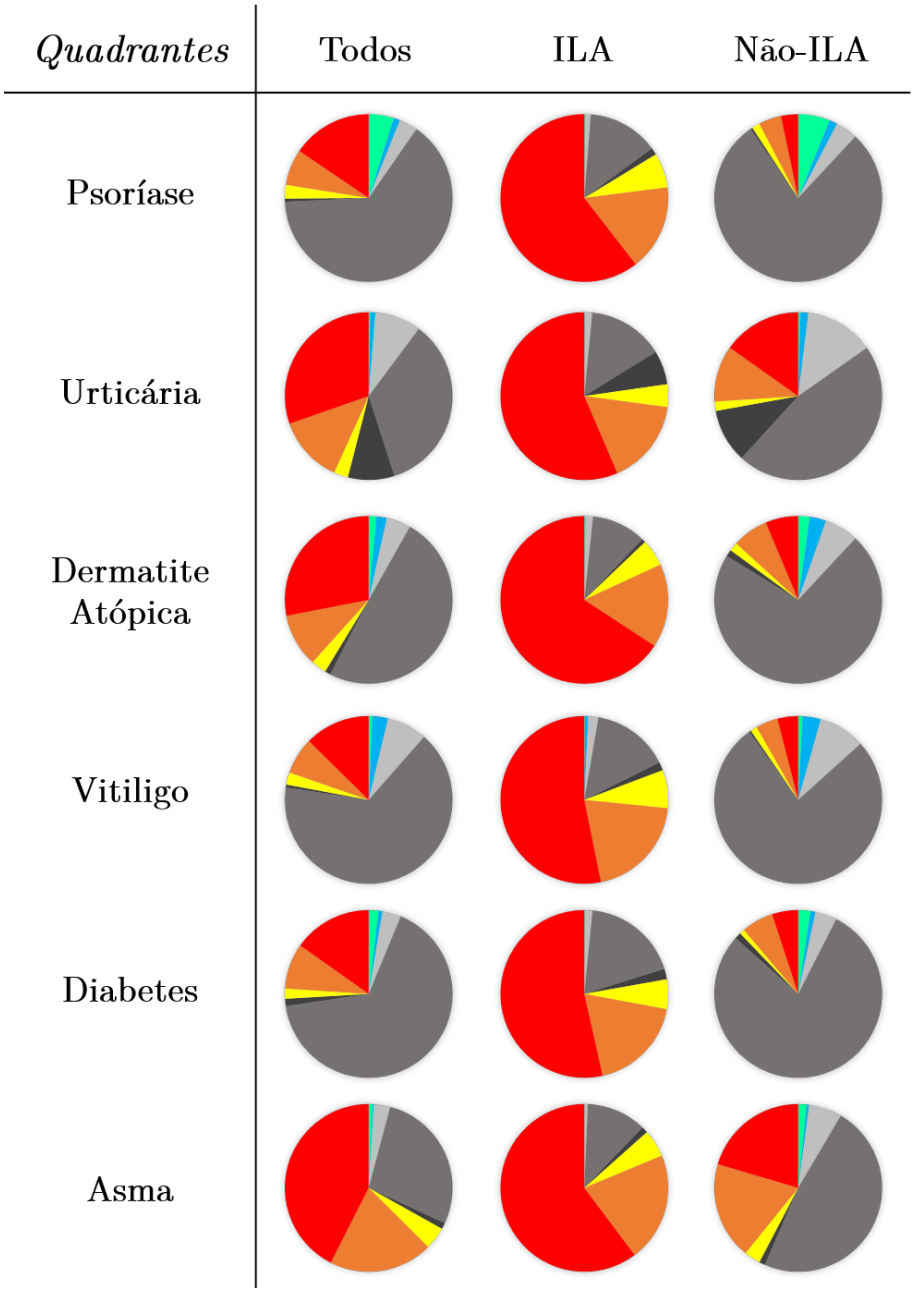

Figura 4.1: Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino exclusivo com dados obtidos via **crowdsourcing** e filtragem realizada tendo em conta o **classificador binário (ILA vs não-ILA)**.

4.3.3 Distribuição dos tweets no tempo

Na presente subsecção são apresentados dois gráficos por doença, cujos eixos verticais marcam o **número de tweets por mês**. O primeiro apresenta a sobreposição das curvas de distribuição ao longo do tempo, do conjunto total dos tweets que mencionam a doença e do conjunto filtrado (ILA). Pretende-se com isto perceber a eficácia da modalidade de filtragem, a partir de alguns eventos extraordinários que acontecem na linha de tempo.

O segundo gráfico, divide o subconjunto ILA respetivo à doença pelos vários agrupamentos emocionais e os dispõe ao longo do tempo. O objetivo passa por perceber a evolução e identificar períodos característicos. Os agrupamentos de valência neutra, representados anteriormente por escalas se cinzentos, são aqui reunidos numa única curva a preto, de forma a melhorar a percetibilidade das várias curvas, sem com isto se perder informação, já que, como mencionado anteriormente, não é relevante subdividir a valência neutra por vários níveis de ativação. Afinal, como se exprime um sentimento neutro com diferentes intensidades?

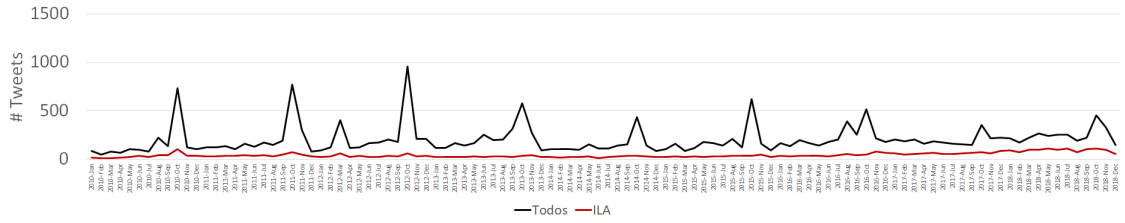

Figura 4.2: Comparação da distribuição do número mensal de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.2: A curva a preto representa a distribuição de todos os tweets que mencionam psoríase ao longo do tempo. A vermelho está representada a distribuição temporal dos tweets ILA, que ignora as muitas publicações de grupos pelo Dia Mundial da Psoríase com sucesso.

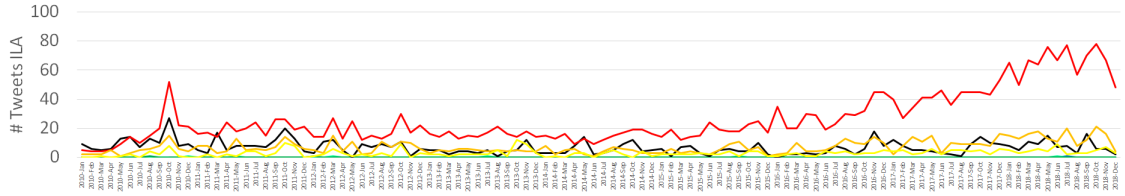

Figura 4.3: Comparação da distribuição do número mensal de tweets que mencionam **psoríase**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.3: Neste gráfico é possível observar que um crescimento da curva a vermelho (V-A+) desde o início de 2017, chegando à publicação de 80 tweets por mês, enquanto as restantes emoções não ultrapassam as 20. Agora com maior resolução, observa-se um pico em Outubro de 2010, provavelmente associado ao Dia Mundial da Psoríase, o que indica a não total resiliência a este evento, por parte do filtro de tweets.

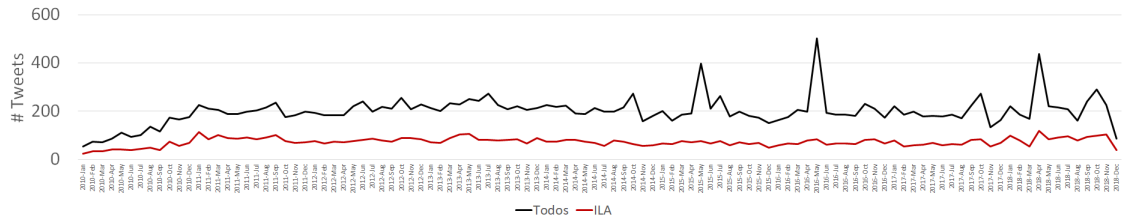

Figura 4.4: Comparação da distribuição do número mensal de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.4: A curva vermelha (ILA) parece acompanhar a tendência da curva a preto durante todo o ano de 2018.

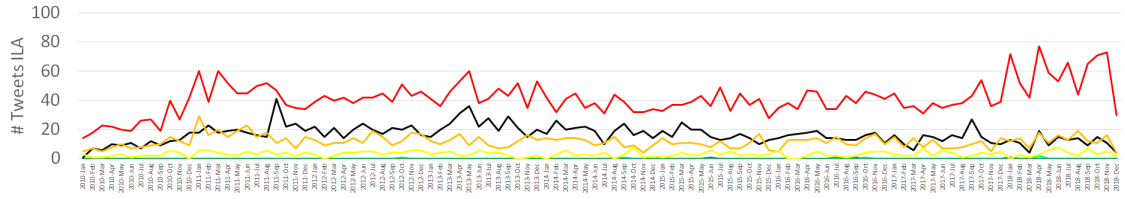

Figura 4.5: Comparação da distribuição do número mensal de tweets que mencionam **urticária**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.5: Não se observa uma tendência suficientemente estável na curva a vermelho.

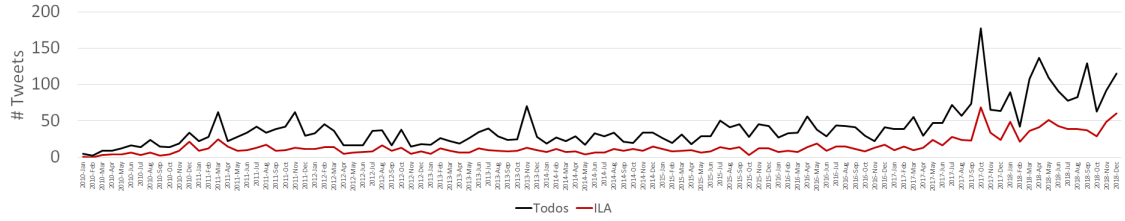

Figura 4.6: Comparação da distribuição do número mensal de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.6: O pico associado ao programa *"Bem Estar"*, parece manter-se na curva ILA. Tal se pode explicar porque as pessoas foram motivas a partilhar o seu testemunho sobre dermatite atópica.
Segue-se um exemplo:

```
Tenho dermatite atópica e sofri muito bulling na época da escola por
os colegas achar q iriam ‘‘pegar minhas perebas" 😔 #vocenobemestar
```

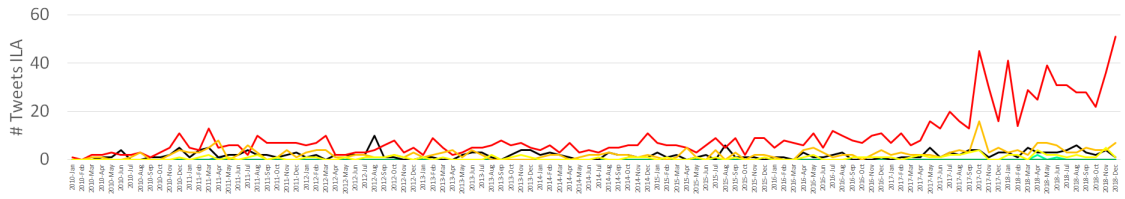

Figura 4.7: Comparação da distribuição do número mensal de tweets que mencionam **dermatite atópica**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.7: A partir de Setembro de 2017, surgem picos no número de publicações associadas às emoções negativas de alta intensidade (V-A+).

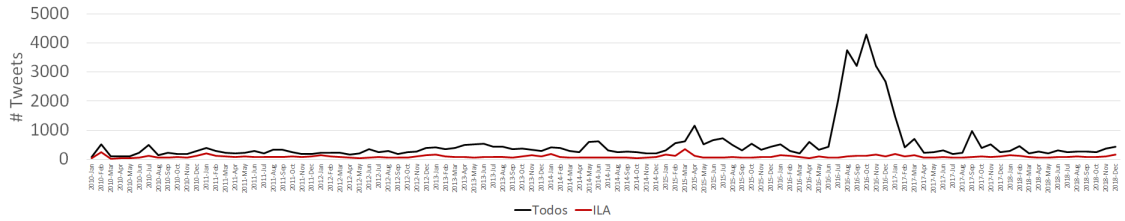

Figura 4.8: Comparação da distribuição do número mensal de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.8: Podemos concluir que o classificador ignorou com sucesso o episódio de publicidade abusiva.

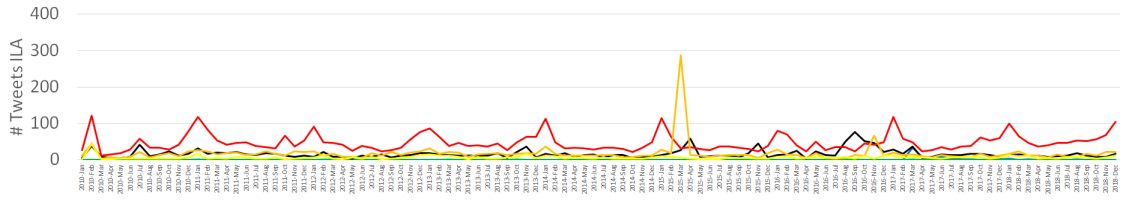

Figura 4.9: Comparação da distribuição do número mensal de tweets que mencionam **vitiligo**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.9: Duas observações podem ser feitas com base neste gráfico. A primeira, é de que ocorre um pico de 300 tweets no mês de Março de 2015, classificado com emoção negativa de média intensidade (V-A0). Uma inspeção ao conjunto de dados permitiu verificar que 270 destes 300 tweets são semelhantes a:

```
na verdade sou negra e tenho vitiligo #VoteRuRushers #KCA
```

Dois erros são cometidos: em primeiro lugar, estes tweets constituem ruído, uma vez que citam o pronunciamento de uma personalidade ligada a uma banda de música, que

faz uso do sentido figurado; e em segundo lugar, não transmite uma emoção negativa, sugerindo que a palavra 'negra' esteja associada a uma emoção negativa, o que não é admissível. Assim se deduz que ambas as tarefas, filtragem e caracterização emocional, falharam neste ponto.
A segunda observação que pode ser feita neste gráfico, é que existe sazonalidade na curva vermelha, alcançando 100 tweets nos meses de Janeiro. No mês de Janeiro é verão no Brasil e uma inspeção aos dados revelou a existência de muitos tweets onde o autor compara a renovação da pele decorrente de uma queimadura solar com vitiligo. Por exemplo:

```
parece que eu tenho vitiligo de tanto q to descascando pqp
```

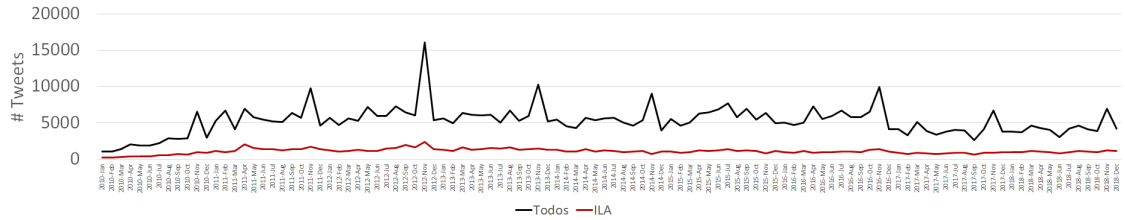

Figura 4.10: Comparação da distribuição do número mensal de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.10: 4.11 com melhor resolução e prevendo a soma das várias curvas, que a filtragem foi bem sucedida, ignorando os picos de Novembro.

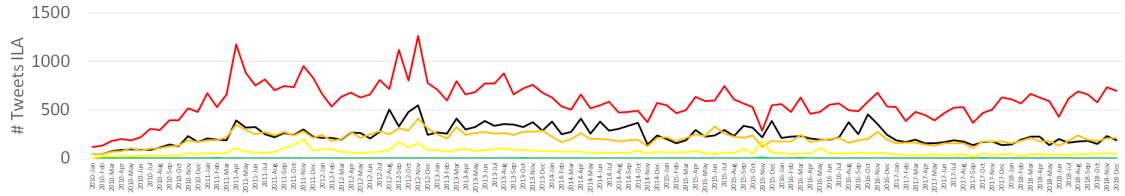

Figura 4.11: Comparação da distribuição do número mensal de tweets que mencionam **diabetes**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.11: Desde 2014, o curva a vermelho decorre com certa estabilidade dentro da janela de 300 a 700 tweets por mês.

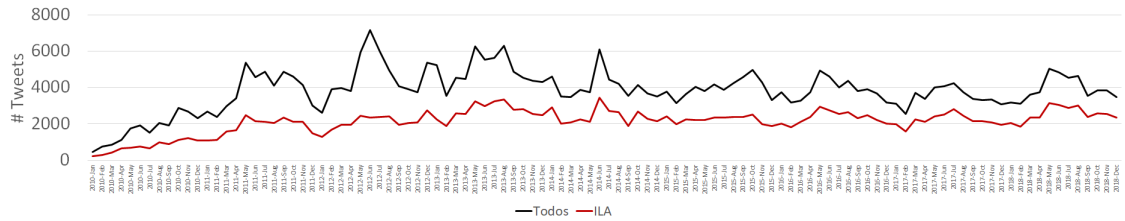

Figura 4.12: Comparação da distribuição do número mensal de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.12: A curva vermelha comporta-se como uma réplica de menor amplitude da curva a negro.

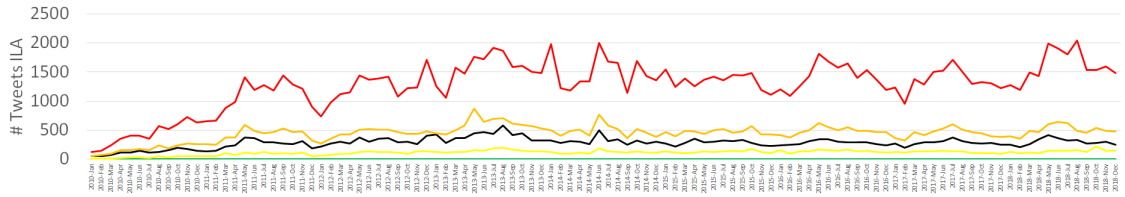

Figura 4.13: Comparação da distribuição do número mensal de tweets que mencionam **asma**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.13: A curva vermelha, aqui representativa da emoção negativa de alta intensidade, parece conservar certa oscilação e zonas baixas dentro daqueles meses em que é verão no Brasil: Dezembro a Março. Ou seja, pela altura do verão existem menos emoções negativas associadas à menção da doença de asma. Fica então a questão se existirá prevalência de asma no inverno.
Num artigo publicado num site de um hospital brasileiro [3] consta que *"com a chegada do inverno crescem o número de casos de asma"*, porque com o frio, que por si só já desencadeia crises de asma, há uma *"maior convivência em locais fechados"*, o que *"favorece a transmissão de doenças infecciosas, sobretudo virais, que também favorecem as crises de asma"*. Pode-se ainda confirmar num estudo brasileiro sobre a prevalência de asma: *"A sazonalidade dos atendimentos por asma, com maior freqüência nos meses de maio, junho e julho (...)"* (Ezequiel et al. (2007)).

A análise emocional ao longo do tempo revela aqui um ponto forte, permitindo correlacionar, até certo ponto, zonas altas e baixas da curva de emoção negativa de alta intensidade com episódios de crise.

Pode ler-se num tweet datado em Junho de 2018:

```
Eu amo inverno, mas quando ataca minha asma já quero o verão de novo
```

Vale lembrar que existem vários tipos de asma, e casos em que o verão também complica a vida dos doentes. Os resultados da análise emocional parecem ir ao encontro dos estudos de prevalência de asma no Brasil.

4.4 Resultados da filtragem com três classificadores

As experiências descritas na secção anterior foram repetidas, com a diferença na filtragem dos tweets, aqui feita pelo cruzamento de resultados de três classificadores individuais: *Autor*, *Sentido* e *Doente*.

4.4.1 Análise de Tópicos

A distribuição dos tópicos consta na tabela 4.13.

[3] `http://www.hospitalmoinhos.org.br/saude-e-voce/a-asma-e-o-inverno/`

#	Tópico
0	eu tenho não psoriase vitiligo URLREF dermatite_atopica mais muito minha sou anos posso sobre descobri doença me bem agora cura
1	urticaria me gente causa pessoas URLREF chega 😇 amor la vejo coisas D= nome palavra ouvir deixa fala tanta ouço
2	vitiligo eu não era ela você meu vi me minha ele tava aí branco URLREF sua mais mim mulher negro
3	dia psoriase 😰 nervosa hoje urticaria saco mundial nacional final novamente atacando 29 outubro louca 😷 combate saúde português evitar
4	urticaria eu tive dias 3 2 1 remédio psoriase hoje começa meu sai tomar uns fui ontem socorro anos fiquei
5	😂 urticaria provoca el con me ninguém usava peruca confirma mi música merece voz lo URLREF autópsia necrópsia filha li
6	bom será dia 😍 dr vitiligo mulheres humor feliz droga sexta feira ❤ homens ¿ pacientes centro marcas curso filme
7	vitiligo não michael ele jackson doença URLREF branco pessoas pele dele era mais filho preconceito mj contagiosa cor mundo sobre
8	urticaria tanta logo quer Twitter odio alguem pensa curiosidade gastrite muita :O moral agonia estou podia cuidado morrendo visto lhe
9	vitiligo :D kkkkk amigos diz le vermelho nesse kkkk chegou azul urticaria limao verão verde entra marta URLREF 🤔
10	psoriase não eu vitiligo nao me outro doença estresse causa passando ligo sentindo auto minha galera oi dia tu pomada
11	urticaria não dermatite_atopica mais eu odeio :(alergia crise psoriase aguento calor frio boa noite merda dormir coceira tengo estou
12	terapia vitiligo 3773 casais 8771 avaliacao 9185 depressao #setelagoas sete lagoas setelagoas is corretivo vamo #euacredito brown mala estressante mc
13	vitiligo eh linda branca mais modelo igual mina maquiagem pescoço tão pegou bonito foto negra hahahaha 👌 família boca homem
14	psoriase minha meu urticaria eu to não corpo novo tô atacada mãe agora cabeça semana dor mais deus voltou cheia
15	minha psoriase vida ataca dermatite_atopica voltando inferno voltar cu não porra dermatologista meu foda emocional caralho atacou preciso pele acabar
16	urticaria 😟 :/ encontro começou paz 50 portadores creme psoriase segunda cedo 6 aos geral sofrendo pânico paulo 1 td
17	pele doenças vitiligo psoriase podem dermatite_atopica mais crônica etc brasil tratamento doença certas câncer povo tv pena outros palavras
18	urticaria me dando la puta produce tu 😒 esa dio pouco daqui aquagênica anti água facebook 😞 alérgico adoro 😩
19	vitiligo eu to parece tenho meu descascando acho tô parecendo cara minha mancha rosto sol estou agora braço tão mano

Tabela 4.13: 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelos três classificadores **(Autor, Sentido, Doente)** treinados com os dados de **crowdsourcing**.

O tópico 0, à semelhança do tópico 0 presente na tabela 4.12, é rico em pronomes e formas verbais que apontam para a exposição de informação do autor sobre ele mesmo.

O tópico 1 é mais um caso de referência a urticária no sentido figurado, expressando a irritação despertada por algumas pessoas no autor ('me','gente'/'pessoas','causa'). O tópico contém um emoticon de conotação negativa ('D=') e um emoji de conotação positiva (sorriso) provavelmente associado à palavra 'amor', tratando-se de ruído dentro do tópico.

Do tópico 2 pode ser extraído um par de termos relacionados a vitiligo ('branco','negro'), e de resto sobram um conjunto variado de pronomes.

O tópico 3 trata do dia mundial da psoríase, presente nos tópicos 8 da tabela 3.3 e 7 da tabela 4.12. Urticária é também mencionada, bem como um emoji de conotação negativa (choro) o termo 'nervosa'.

Do tópico 4 sobressai o termo 'socorro', que sugere a expressão de urgência/gravidade com que o autor pede ajuda para o tratamento da dermatose, e a ocorrência dos termos 'remédio' e 'tomar'.

Já do tópico 5 sobressai a presença de termos em língua espanhola ('el','con', 'mi', 'lo') e termos associados à notícia das descobertas feitas na inspeção ao corpo de Michael

Jackson ('confirma', 'peruca', 'autópsia', 'necrópsia'), não mencionando no entanto o nome do cantor.

O tópico 6 inclui bastantes sinais de positividade ('bom', 'humor', 'feliz', '😍', '❤️').

o tópico 7 menciona o vitiligo de Michael Jackson, e a referência a 'filho' pode indicar a notícia de que um filho do artista também seria portador da condição, como retratado no tópico 4 da tabela 3.3.

O tópico 8 parece seguir o mesmo sentido do tópico 1, na referência a urticária no sentido figurado. Pesquisando sobre a possível associação entre urticária e gastrite, se toma conhecimento de um possível agente infeccioso comum, Helicobacter pylori (Criado et al. (1999)).

O tópico 9 parece reunir tweets semelhantes a estes:

```
tá parecendo que eu tenho vitiligo vermelho =/
```

```
eu to ficando louco ou o meu vitiligo ta ficando verde?  =x
```

```
referencias rato?azul e vermelho será que faz relaçao com o vitiligo
USERREF
```

Este último está classificado como 'não-ILA' pelo classificador binário. Os dois primeiros exemplos estão classificados como 'ILA' por ambas as abordagens.

O tópico 10 menciona 'estresse', 'pomada' e as doenças vitiligo e psoríase. Tem pronomes ('eu', 'me') que apontam na referência própria do autor, mas outros sinais dissidentes ('tu', 'outro').

O tópico 11 vai ao encontro de outros tópicos presentes nas tabelas anteriores, que fazem referência à urticária colinérgica e de frio. Também sugere o relato de sintomas e agravamento ('coceira', 'crise', 'não' 'aguento' / 'não' 'dormir'). Este tópico está emocionalmente carregado ('odeio', ':('.'merda').

O tópico 12 reflete o episódio da publicidade abusiva a acompanhamento psicológico, também disponível a doentes de vitiligo.

O tópico 13 expressa comentários positivos ('linda', 'bonito', '👌') sobre a aceitação de vitiligo, com provável referência à 'modelo' Winnie Harlow.

O tópico 14 é bastante heterogéneo. Destaca-se a referência a 'corpo', 'cabeça' e 'dor', como possíveis indicadores da extensão e gravidade dos sintomas.

O tópico 15 contém bastantes vocábulos do calão ('cu', 'porra', 'foda', 'caralho') e referência a 'inferno', indicadores óbvios de um estado 'emocional' negativo, relatado pelo próprio sujeito.

O tópico 16 parece misturar vários assuntos, como o relato pessoal de sofredores de uma doença, caracterizado pelo emoji e emoticon, e provavelmente um 'encontro' de médicos dermatologistas na cidade de São 'Paulo' para debater psoríase e vitiligo, que no presente ano de 2019 teve a sua 17ª edição.

O tópico 17 é bastante geral e traduz provavelmente um carácter informativo ('tv', 'povo') já que menciona três dermatoses e ainda cancro ('câncer').

O tópico 18 contém palavras em língua espanhola (‘la’,‘produce’,‘esa’,‘dio’) como outros tópicos em tabelas anteriores. Inclui três emojis de expressão negativa, contrariados pelo verbo adorar (‘adoro’). Tratando-se de um tópico exclusivo à doença de urticária, encontra em três termos sucessivos a definição de um dos tipos: ‘aquagénica anti água’, evidenciando o poder de extração de conhecimento da técnica LDA.

O tópico 19 traduz a comparação muito de comum da aparência da recuperação da pele após queimadura solar com vitiligo.

Apreciação

A tabela 4.13 inclui 6 tópicos com o termo ‘`URLREF`’, um a mais que a tabela 4.12, não sendo ainda o suficiente para concluir que a filtragem a partir do classificador binário seja mais eficaz, porque a diferença de uma unidade não traduz uma diferença considerável, tendo em conta a volatilidade da técnica LDA na obtenção dos tópicos.
Já o número de emojis é reduzido para 14 e o de emoticons é 4. A tabela 4.12 contém 22 elementos gráficos, mais 4 do que a presente.
Pela análise especulativa, admitem-se **10** tópicos compatíveis com tweets tweets ILA (0, 3, 4, 9, 10, 11, 14, 15, 16, 18), sendo que **3** deles (3, 9, e 16), têm um nível alto de ambiguidade ou mistura de assuntos.

4.4.2 Emoções

A figura 4.14 apresenta os resultados da classificação de emoção, com os mesmo classificadores de valência e ativação, mas com um conjunto de dados filtrados algo diferente.

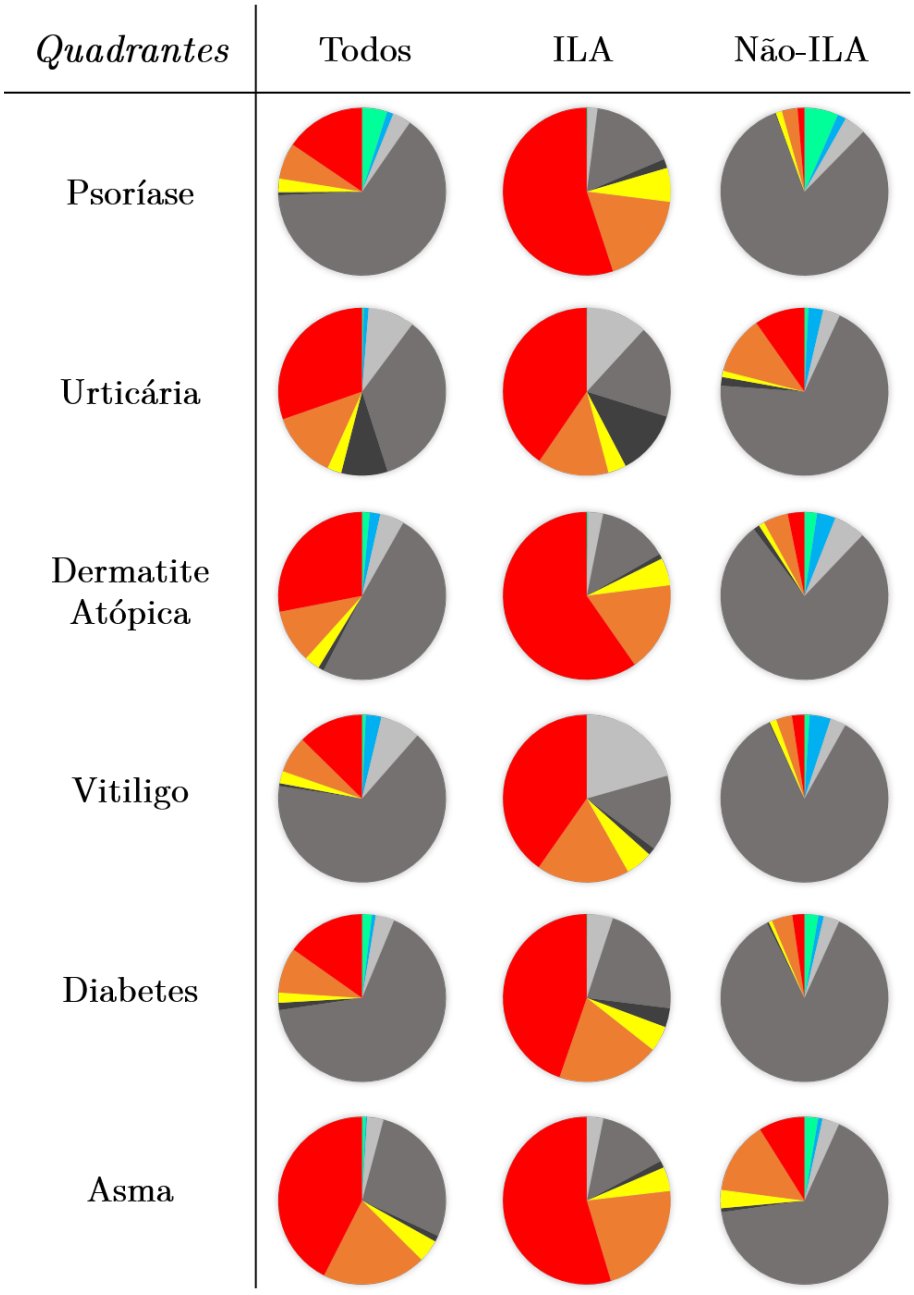

Figura 4.14: Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino exclusivo com dados obtidos via **crowdsourcing** e filtragem realizada tendo em conta os **três classificadores (Autor, Sentido, Doente)**.

De forma semelhante aos gráficos presentes na figura 4.1, a figura 4.14 mostra a ausência das cores que representam as emoções de valência positiva nos subconjuntos ILA. Mais uma vez, diabetes e vitiligo apresentam uma semelhança considerável nos três gráficos. No que toca aos subconjuntos ILA, parece admissível agrupar as doenças em duas instâncias:

- Psoríase, Dermatite Atópica e Asma
- Urticária, Vitiligo e Diabetes

Segue-se a ordenação decrescente das percentagens dos grupos emocionais de valência negativa dos subconjuntos ILA:

V-A+ (vermelho): dermatite atópica (59.7%), psoríase (55.1%), asma (54.7%), diabetes (44.7%), urticária (40.4%), vitiligo (40.3%).

V-A0 (laranja): asma (22.2%), diabetes (19.7%), psoríase (18.0%), vitiligo (17.9%), dermatite atópica (17.3%), urticária (13.7%).

V-A- (amarelo): psoríase (6.5%), dermatite atópica (5.3%), vitiligo (5.2%), diabetes (4.9%), asma (4.7%), urticária (3.5%).

4.4.3 Distribuição dos tweets no tempo

Nesta etapa de observação dos gráficos, foram procuradas diferenças relativamente aos resultados anteriores. Note-se que a linha vermelha nos gráficos 'Todos vs ILA' não corresponde ao tipo de emoção, mas a todos os tweets identificados como ILA.

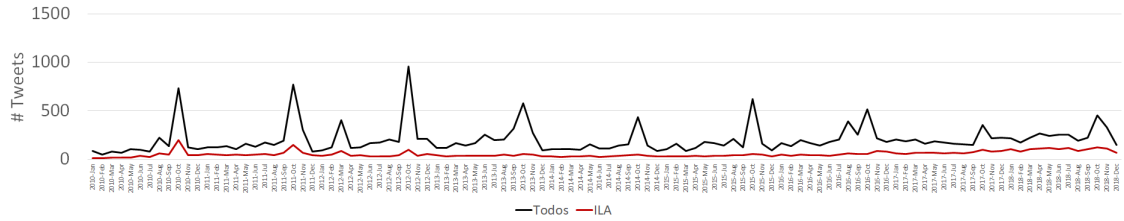

Figura 4.15: Comparação da distribuição de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.15: Observa-se uma menor capacidade de filtragem no primeiro pico associado ao Dia Mundial da Psoríase, relativamente à modalidade anterior, já que a curva vermelha da figura 4.2 apresenta um vértice menos acentuado em Outubro de 2010 e 2011.

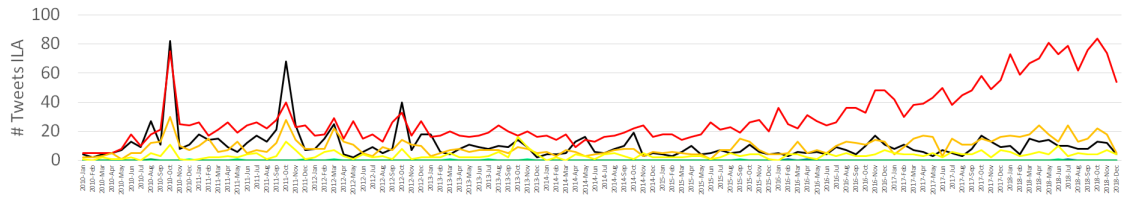

Figura 4.16: Comparação da distribuição de tweets que mencionam **psoríase**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.16: Como apontado anteriormente, esta modalidade de filtragem apresentou um ponto negativo, não filtrando tweets publicados por organizações

que nem sequer citam doentes. É possível dizer com alguma razoabilidade que os classificadores de emoção estão a caracterizar aproximadamente metade destes tweets de forma correta (neutra (V0)), a preto, e outra metade incorretamente, a vermelho (V-A+). Vemos isso na sobreposição das linhas vermelha e preta no primeiro pico notável, que já sabemos se tratar do dia mundial da psoríase. No entanto, esta consideração não é infalível, porque apesar de ser uma época de muitas publicações por parte de organizações, os doentes também são motivados a partilhar o seu testemunho. Em consenso com os resultados da modalidade abordada anteriormente, podemos concluir o crescimento consistente da curva representativa de emoções negativas de alta intensidade.

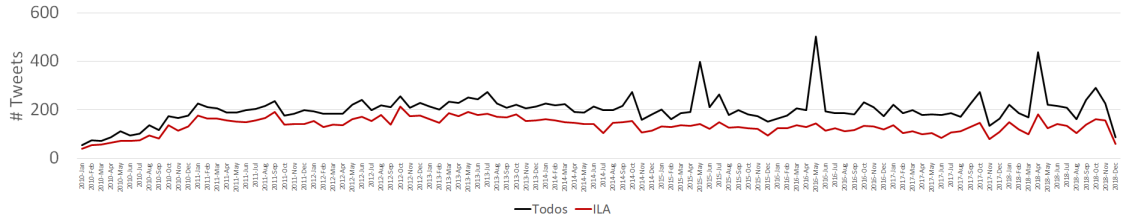

Figura 4.17: Comparação da distribuição de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.17: A curva 'ILA' apresenta-se mais "colada" à curva 'Todos', do que aquela obtida na figura 4.4.

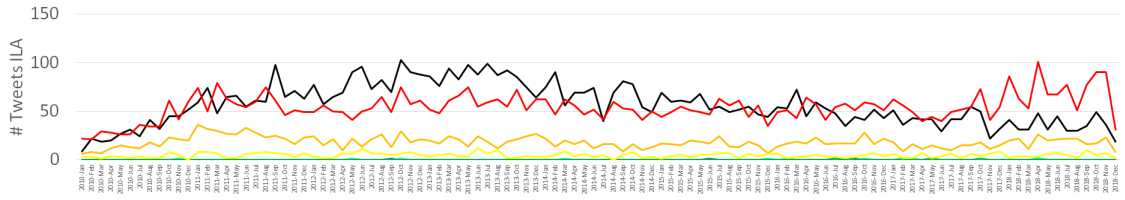

Figura 4.18: Comparação da distribuição de tweets que mencionam **urticária**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.18: Este gráfico apresenta a curva a vermelho muito semelhante em forma em relação à curva a vermelho do gráfico 4.5, com a amplitude a chegar a mais 20 publicações por mês. A curva correspondente a tweets neutros é aqui elevada à mesma relevância da curva vermelha.

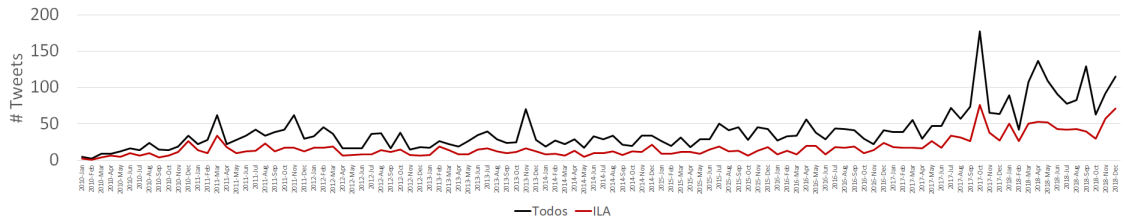

Figura 4.19: Comparação da distribuição de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.19: Não existem diferenças significativas entre esta figura e a figura 4.6.

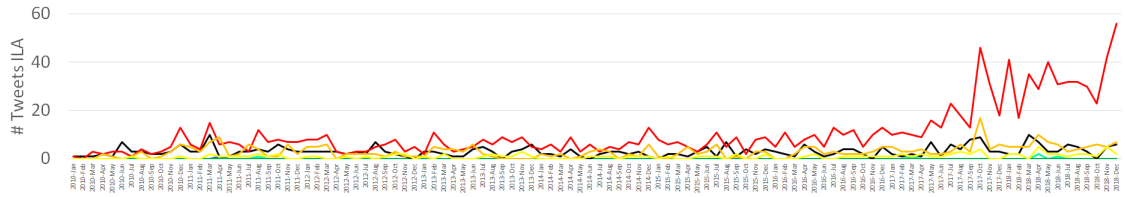

Figura 4.20: Comparação da distribuição de tweets que mencionam **dermatite atópica**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.20: De forma semelhante à figura anterior, esta apresenta-se equivalente à obtida para a filtragem com *classificador binário*.

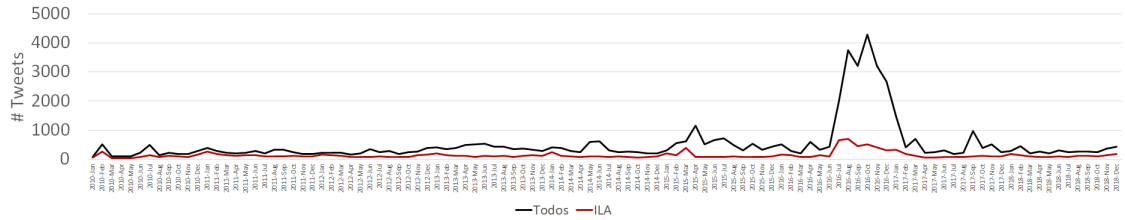

Figura 4.21: Comparação da distribuição de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.21: Daqui se deduz realmente um ponto negativo. Enquanto na distribuição relativa a psoríase, a curva ILA segue os primeiros picos do dia comemorativo, podendo tratar-se realmente da partilha de testemunhos nessa época de sensibilização, aqui a distorção é causada por um episódio abusivo de publicidade, que deveria ser totalmente rejeitado à semelhança do que é visto na 4.8.

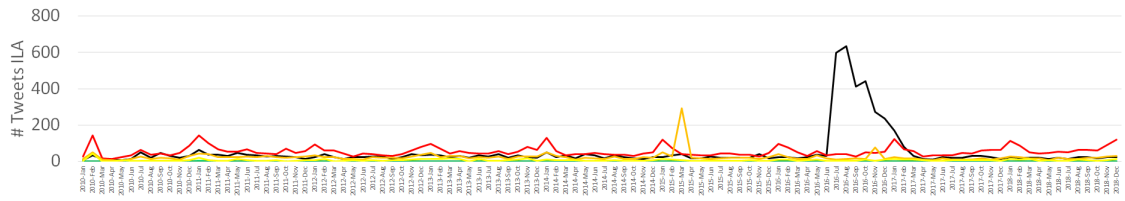

Figura 4.22: Comparação da distribuição de tweets que mencionam **vitiligo**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.22: Este gráfico comprova, com maior resolução, as considerações anteriores. No entanto destaca-se um ponto positivo: o classificador de emoção, concretamente o de valência, já que a ativação não está a ser avaliada nos tweets neutros, catalogou estes milhares de tweet de forma correta.

Comentário - Figura 4.23: Não se verificam diferenças significativas ao resultado

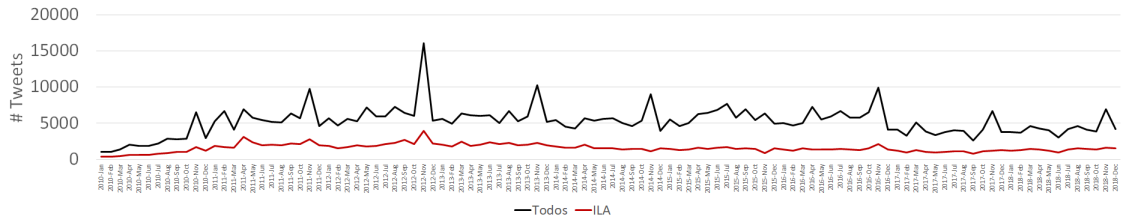

Figura 4.23: Comparação da distribuição de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

obtido na modalidade anterior, apenas a maior contabilização de tweets, e a acentuação dos vértices nos meses de Abril de 2011 e Novembro de 2012.

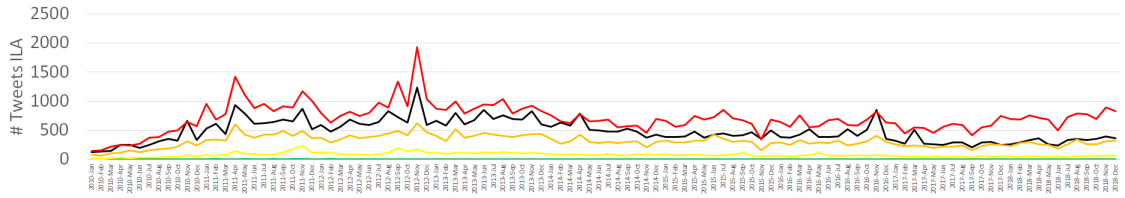

Figura 4.24: Comparação da distribuição de tweets que mencionam **diabetes**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.24: Mais uma vez os resultados são semelhantes à modalidade anterior (classificador binário), com a exceção de que a curva dos tweets neutros (a preto), foi elevada e situada entre a emoção negativa de média intensidade (V-A0/laranja) e a emoção negativa de alta intensidade, confirmando-se que os tweets adicionados são de carácter emocional neutro.

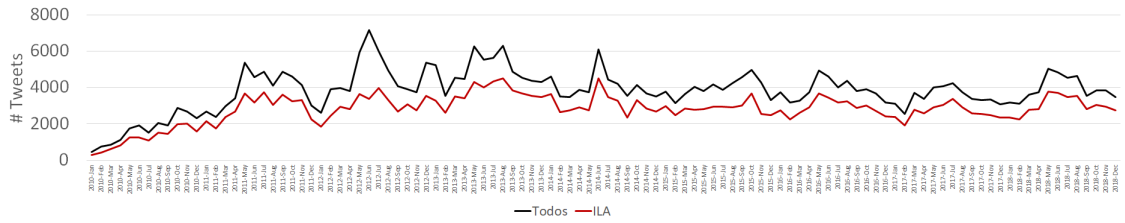

Figura 4.25: Comparação da distribuição de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

Comentário - Figura 4.25: Nesta modalidade, ouve um menor número de tweets a serem considerados ruído.

Comentário - Figura 4.26: Neste gráfico, denota-se um acréscimo na amplitude da curva vermelha, evidenciando a diferenças das zonas altas e das zonas baixa, que vão ao encontro da prevalência de Asma no Brasil, durante o inverno.

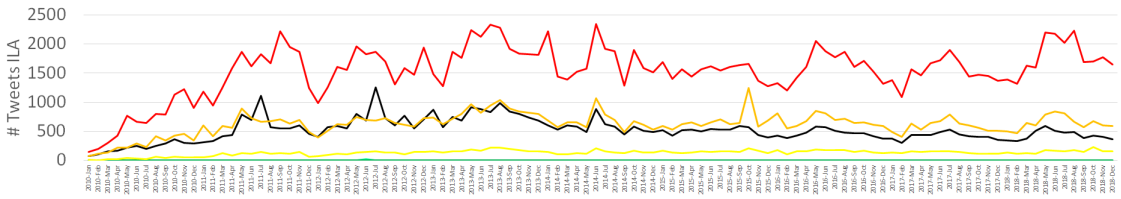

Figura 4.26: Comparação da distribuição de tweets que mencionam **asma**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**.

4.5 Discussão

Neste capítulo foram comparadas duas modalidades de filtragem de tweets ILA: o *Classificador Binário* e o cruzamento de resultados dos classificadores de *Autor*, *Sentido* e *Doente.*

Tecem-se agora algumas considerações sobre qual a mais adequada:

1. Dos valores de performance listados na tabela 4.8, não se pode prever qual a melhor modalidade, porque não estão a ser validados na mesma tarefa, embora estejam relacionadas. O mais apropriado seria obter métricas de performance no cruzamento de resultados dos três classificadores ASD e aí comparar com o *Classificador Binário.* Tal não foi possível por dificuldades na adaptação das funções de *cross-validation* já disponibilizadas pela biblioteca *Scikit-Learn.*

2. Da comparação dos gráficos circulares também nada se pode concluir.

3. Quanto à análise de tópicos, embora no caso do classificador binário se tenham considerado apenas 8 tópicos representativos de tweets ILA, os tópicos resultantes da segunda modalidade são mais ambíguos e confusos, de tal forma que é mais prudente considerar apenas 7, não constituindo ainda um critério definitivo.

4. O classificador binário apresenta assim, melhores indicadores do que a modalidade alternativa de filtragem, no que é percetível na análise de tópicos: um número menor de tópicos contendo o termo '`URLREF`' (indiciador de tweet não-ILA), maior presença de emojis/emoticons (22 contra 18), e o maior número de tópicos passíveis de representar tweets ILA (8 contra 7).

5. Da distribuição no tempo, é confirmada a preferência no classificador binário, por ter sido resiliente no episódio de publicidade abusiva.

Contudo, surge a necessidade da procura por uma solução mais robusta, especialmente para melhorar as classificações de valência e ativação, cuja performance de validação é muito baixa.

O número de tópicos compatíveis com tweets ILA não ultrapassa a metade, evidenciando também a necessidade de obter um mecanismo de filtragem mais eficiente.

Uma das limitações dos resultados obtidos será a quantidade de dados anotados, bem como a sua distribuição por doenças. Relembramos que uma grande parte dos tweets anotados mencionam dermatite atópica e, por outro lado, nenhum desses tweets menciona exclusivamente vitiligo, asma ou diabetes. Na expectativa de que uma maior quantidade de

dados anotados possa melhorar os resultados e as análises realizadas, decidimos aumentar o conjunto de dados anotados. Para tal, a nossa primeira opção seria voltar a recorrer à plataforma Figure-Eight. No entanto, isto não foi possível porque a plataforma esteve várias semanas sem funcionar corretamente, o que se terá devido a problemas de gestão e a uma mudança de gerência. Com o aproximar do final do trabalho e na impossibilidade de recorrer a crowdsourcing, tivemos de considerar opções alternativas. A opção pela anotação de um novo conjunto de tweets pelo autor da tese. De forma a optimizar os classificadores, os tweets anotados foram selecionados através de um processo de Active Learning. Os resultados desta abordagem são descritos no próximo capítulo.

Capítulo 5

Classificação de tweets com base em Active Learning

No presente capítulo é descrita a implementação em Active Learning, adoptada para reunir um maior número de tweets anotados. As mesmas experiências do capítulo anterior são aqui repetidas.

5.1 Implementando Active Learning

Devido a problemas na plataforma **Figure-Eight**[1], não foi possível obter o número desejado de tweets anotados. Na revisão de métodos usados para análise de sentimento de tweets, que tratam do estado de saúde, relaizada por Gohil et al. (2018), consta uma pequena tabela de estudos de classificação, usando KNN, Naïve Bayes e SVM, bem como o tamanho de alguns corpus anotados: 250, 500, 1000, 2200 e 3000. Os estudos abordam diferentes tópicos e comportam abordagens distintas, pelo que não é possível deduzir a melhor combinação de decisões, já que grande parte da tabela está incompleta. Havendo corpus com tamanho superior a 1000, era objectivo anotar, pelo menos, mais algumas centenas. Os problemas na plataforma não viram solução, e em alternativa, por restrição de tempo, foi decidido acrescentar anotações próprias em vez de recorrer a outra plataforma online de anotações.

Foi realizada alguma pesquisa por métodos para realizar uma anotação mais inteligente, i.e., seleccionar à partida o conjunto de tweets a anotar, não aleatoriamente, que de alguma forma tenderiam a representar melhor, pela informação que oferecem ao conjunto anotado, o universo dos tweets obtidos. Não encontrando resposta, foram realizadas algumas experiências, como o cálculo da semelhança entre tweets ou o agrupamento usando técnicas de clustering. Ainda não obtidos resultados, o Professor Orientador sugeriu o recurso a *Active Learning*.

As primeiras experiências não decorreram de uma implementação correcta, anotando-se tweets que não correspondiam aos sugeridos para anotação, mas uma vez que foram anotados, não foram depois descartados. Era também de interesse contemplar as doenças que ainda não tinham tweets anotados. Esta abordagem de *Active Learning* requer a anotação de um só tweet de cada vez, seguida de novo cálculo e sugestão de novo tweet para anotar. Tendo em conta que se tratava de um processo muito demoroso, da necessi-

[1] `https://www.figure-eight.com/`

dade de beneficiar as diferentes tarefas de classificação e ainda contemplar especificamente algumas doenças, a anotação foi feita da seguinte maneira: em vez de ser obtido um tweet de cada vez para anotação, numa primeira fase foi obtido um tweet do conjunto de vitiligo, um do conjunto de diabetes e outro do conjunto de asma (conjuntos sem tweets anotados), para cada uma das tarefas de classificação (Autor, Sentido, Doente, Valência e Ativação). Assim foi feito durante algumas iterações, adicionando 15 novos tweets ao conjunto anotado (3 doenças × 5 classificações) de cada vez. Posteriormente, os subconjuntos não anotados das 6 doenças foram reunidos, de forma a priorizar a obtenção de anotações mais determinantes, cinco a cada iteração (1 por classificação), apostando na capacidade de generalização entre doenças, como sugerido por Yin et al. (2015), em prol do ganho de informação.

A *framework* utilizada para a implementação da estratégia de *Active Learning* tem por nome **modAL** [2], desenvolvida por Tivadar Danka e Peter Horvath.

Estabilizada a implementação em *Active Learning*, foi adoptado o algoritmo `MultinomialNB` como classificador e *estratégia de query* de incerteza, classificação menos confiante, que devolve o tweet com o maior valor de incerteza da previsão. Segue um exemplo ilustrativo na tabela 5.1, que mostra qual de três tweets deve ser recomendado para anotação, segundo o critério de classificação menos confiante.

Dentro do conjunto de critérios de incerteza, existem mais duas variantes: a margem de classificação, calculada subtraindo a segunda maior probabilidade à primeira (para o caso do tweet B: $M_B = 0.6 - 0.3$), e a entropia ($H_B = -0.6\,ln(0.6) - 0.3\,ln(0.3) - 0.1\,ln(0.1)$).

Tweet	**Probabilidades/Doente**			**Classificação**	**Incerteza**
	Autor	**Outra**	**NA**		
A	0.1	0.85	0.05	Outra	$1 - 0.85 = 0.15$
B	0.6	0.3	0.1	Autor	$1 - 0.6 = 0.4$
C	0.39	0.61	0.0	Outra	$1 - 0.61 = 0.39$

Tabela 5.1: Active Learning - exemplo ilustrativo: cálculo da incerteza (menor confiança na classificação), critério para recomendar instância para anotação. Neste caso, o Tweet B seria o escolhido.

Foram utilizadas aqui, naturalmente, as mesmas labels para anotar os tweets. Para valência e ativação foi utilizada a escala original de 5 valores.

5.2 Resultados da anotação

Nas tabelas 5.2 e 5.3, podemos destacar a muito baixa proporção de tweets ILA entre os tweets anotados referentes a vtiligo e diabetes, que constitui um ponto negativo ao conjunto exclusivo a crowdsourcing.

Na tabela 5.6 podemos verificar o melhor balanceamento entre os três níveis de ativação (simples). A percentagem de tweets anotados com valência positiva (+), teve também um ligeiro aumento face ao conjunto exclusivo ao crowdsourcing.

Na distribuição por quadrantes e fronteiras, vemos que as percentagens totais associadas à valência positiva (V+) tiveram um ligeiro aumento, e as regiões V0A0 e V-A+, que anteriormente correspondiam a 2/3 do conjunto anotado, tiveram a sua proporção diminuída para metade do conjunto.

[2] `https://github.com/modAL-python/modAL`

#	Autor			Sentido		Doente			Binário	
	Indiv	**Grupo**	**NA**	**Lit**	**Fig**	**Autor**	**Outra**	**NA**	**ILA**	**Não-ILA**
Psoríase	72	15	19	102	4	52	7	47	50	56
Urticária	84	5	7	47	49	55	7	34	29	67
D. Atópica	209	64	18	287	4	143	49	99	140	151
Vitiligo	55	26	0	72	9	8	39	34	7	74
Diabetes	86	50	0	89	47	14	13	109	10	126
Asma	73	18	0	76	15	26	28	37	24	67
TOTAL	579	178	44	673	128	298	143	360	260	541

Tabela 5.2: Distribuição dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores.

%	Autor			Sentido		Doente			Binário	
	Indiv	**Grupo**	**NA**	**Lit**	**Fig**	**Autor**	**Outra**	**NA**	**ILA**	**Não-ILA**
Psoríase	67.9	14.2	17.9	96.2	3.8	49.1	6.6	44.3	47.2	52.8
Urticária	87.5	5.2	7.3	49.0	51.0	57.3	7.3	35.4	30.2	69.8
D. Atópica	71.8	22.0	6.2	98.6	1.4	49.1	16.8	34.0	48.1	51.9
Vitiligo	67.9	32.1	0.0	88.9	11.1	9.9	48.1	42.0	8.6	91.4
Diabetes	63.2	36.8	0.0	65.4	34.6	10.3	9.6	80.1	7.4	92.6
Asma	80.2	19.8	0.0	83.5	16.5	28.6	30.8	40.7	26.4	73.6
TOTAL	72.3	22.2	5.5	84.0	16.0	37.2	17.9	44.9	32.5	67.5

Tabela 5.3: Distribuição em **percentagem** dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores.

#	Valência-5					Ativação-5					Val. Simples			Ati. Simples		
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**-**	**0**	**+**	**-**	**0**	**+**
Psoríase	15	29	39	8	3	5	14	35	23	15	52	39	15	25	35	46
Urticária	10	30	52	3	1	6	6	60	18	6	40	52	4	12	60	24
D. Atópica	65	84	116	16	4	5	6	123	101	49	155	116	20	13	123	155
Vitiligo	1	16	50	12	2	23	15	28	12	3	17	50	14	38	28	15
Diabetes	5	38	58	26	9	42	28	28	24	14	43	58	35	70	28	38
Asma	11	29	37	11	3	19	23	27	15	7	40	37	14	42	27	22
TOTAL	107	226	352	76	22	100	92	301	193	94	347	352	102	200	301	300

Tabela 5.4: Distribuição dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelos vários níveis de valência e ativação (1 a 5), e nas classes geradas posteriormente (simplificação para três labels possíveis: Negativo (-), Origem (0), Positivo (+).

%	Valência-5					Ativação-5				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Psoríase	14.2	27.4	36.8	7.5	2.8	4.7	13.2	33.0	21.7	14.2
Urticária	10.4	31.3	54.2	3.1	1.0	6.3	6.3	62.5	18.8	6.3
D. Atópica	22.3	28.9	39.9	5.5	1.4	1.7	2.1	42.3	34.7	16.8
Vitiligo	1.2	19.8	61.7	14.8	2.5	28.4	18.5	34.6	14.8	3.7
Diabetes	3.7	27.9	42.6	19.1	6.6	30.9	20.6	20.6	17.6	10.3
Asma	12.1	31.9	40.7	12.1	3.3	20.9	25.3	29.7	16.5	7.7
TOTAL	13.4	28.2	43.9	9.5	2.7	12.5	11.5	37.6	24.1	11.7

Tabela 5.5: Distribuição em **percentagem** da reunião dos dados anotados via crowdsourcing com as anotações com recurso a **active learning** pelos vários níveis de **valência** e **ativação** (1 a 5).

As mesmas experiências para apurar as melhores combinações de algoritmo e tipo de features por problema, foram repetidas neste capítulo. O algoritmo `NearestCentroid` veio substituir o algoritmo `LinearSVC` no problema do *Sentido* e com ela a mudança de vectorização de contagem absoluta para TF-IDF. No caso das coordenadas emocionais, foi trocada a utilização de acentos entre valência e ativação. O número máximo de features usadas passou de 1498 para 2020.

%	Valência Simples			Ativação Simples		
	-	0	+	-	0	+
Psoríase	49.1	36.8	14.2	23.6	33.0	43.4
Urticária	41.7	54.2	4.2	12.5	62.5	25.0
D. Atópica	53.3	39.9	6.9	4.5	42.3	53.3
Vitiligo	21.0	61.7	17.3	46.9	34.6	18.5
Diabetes	31.6	42.6	25.7	51.5	20.6	27.9
Asma	44.0	40.7	15.4	46.2	29.7	24.2
TOTAL	43.3	43.9	12.7	25.0	37.6	37.5

Tabela 5.6: Distribuição em **percentagem** da reunião dos dados anotados via crowdsourcing com as anotações com recurso a **active learning** pelos vários níveis de **valência** e **ativação** simplificados: Negativo (-), Origem (0), Positivo (+).

#	Referêncial Emocional								
	V+A-	V+A0	V+A+	V0A-	V0A0	V0A+	V-A+	V-A0	V-A-
Psoríase	2	4	9	9	26	4	33	5	14
Urticária	0	3	1	6	40	6	17	17	6
D. Atópica	2	12	6	9	94	13	136	17	2
Vitiligo	2	5	7	28	19	3	5	4	8
Diabetes	2	8	25	45	9	4	9	11	23
Asma	3	5	6	19	12	6	10	10	20
TOTAL	11	37	54	116	200	36	210	64	73

Tabela 5.7: Distribuição dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelos vários **quadrantes** e fronteiras do referêncial emocional adoptado. A identificação do tweet num quadrante foi feita com base nas anotações de valência e ativação (1 a 5).

%	Referêncial Emocional								
	V+A-	V+A0	V+A+	V0A-	V0A0	V0A+	V-A+	V-A0	V-A-
Psoríase	1.9	3.8	8.5	8.5	24.5	3.8	31.1	4.7	13.2
Urticária	0.0	3.1	1.0	6.3	41.7	6.3	17.7	17.7	6.3
D. Atópica	0.7	4.1	2.1	3.1	32.3	4.5	46.7	5.8	0.7
Vitiligo	2.5	6.2	8.6	34.6	23.5	3.7	6.2	4.9	9.9
Diabetes	1.5	5.9	18.4	33.1	6.6	2.9	6.6	8.1	16.9
Asma	3.3	5.5	6.6	20.9	13.2	6.6	11.0	11.0	22.0
TOTAL	1.4	4.6	6.7	14.5	25.0	4.5	26.2	8.0	9.1

Tabela 5.8: Distribuição em **percentagem** dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelos vários **quadrantes** e fronteiras do referêncial emocional adoptado. A identificação do tweet num quadrante foi feita com base nas anotações de valência e ativação (1 a 5).

Verifica-se uma descida de desempenho, segundo a medida F1 (pesada), nos classificadores de *Autor*, *Sentido*, *Doente*, *Valência* e *Ativação* (3 valores). O classificador binário dobrou o desvio padrão na medida F1 (pesada).

Classificação	Modelo	Lemas	N-gramas	Vectorização	Remoção Stop-words	Remoção Acentos	Número Features
Autor	LinearSVC	Sim	1-2	TF	Não	Sim	1902
Sentido	NearestCentroid	Sim	1-2	TF-IDF	Não	Não	1995
Doente	LinearSVC	Não	1-2	TF	Não	Sim	2017
Class. Binário	LinearSVC	Não	1	TF	Não	Sim	1022
Valência	NearestCentroid	Sim	1-2	TF-IDF	Não	Sim	1993
Ativação	NearestCentroid	Não	1-2	TF-IDF	Não	Não	2020
Valência-5	NearestCentroid	Não	1	TF-IDF	Não	Não	1035
Ativação-5	MultinomialNB	Não	1	Count	Não	Não	1035

Tabela 5.9: Classificadores adoptados. Conjunto de treino: *crowdsourcing* e *Active Learning*. *Class. Binário* corresponde ao classificador binário que separa tweets cujo autor é o doente e fala no sentido literal, dos restantes.

Classificação	Accuracy	Precision	Recall	F1	F1 (pesado)
Autor	0.931 ± 0.033	0.910± 0.041	0.902± 0.044	0.904± 0.037	0.931 ± 0.034
Sentido	0.931±0.033	0.910± 0.041	0.902± 0.044	0.904± 0.037	0.881 ± 0.055
Doente	0.790 ±0.079	0.766± 0.101	0.724± 0.075	0.730 ±0.087	0.789 ± 0.078
Class. Binário	0.841 ±0.061	0.776± 0.094	0.747± 0.106	0.748± 0.105	0.834 ± 0.066
Valência	0.704± 0.052	0.662± 0.139	0.581± 0.061	0.590 ±0.071	0.692 ± 0.047
Ativação	0.592± 0.108	0.503± 0.101	0.495± 0.099	0.474± 0.083	0.583 ± 0.112
Valência-5	0.547± 0.069	0.263± 0.060	0.279± 0.056	0.267± 0.056	0.527 ± 0.064
Ativação-5	0.449± 0.132	0.268± 0.081	0.264± 0.070	0.249± 0.063	0.436 ± 0.129

Tabela 5.10: Performance dos classificadores a partir de *10-fold cross-validation*. Precision-macro, Recall-macro, F1-macro e F1-weighted. Conjunto de treino: crowdsourcing e **active learning**.

Nas tabelas 5.11 e 5.12, verifica-se uma diminuição de tweets ILA, em todas as doenças.

Uma vez aplicadas as duas modalidades de classificação para identificar os tweets ILA, verificamos na tabela 5.13 um aumento de correspondência em todas as doenças, face aos resultados obtidos do treino com o conjunto exclusivo aos dados de crowdsourcing.

#	Autor		Sentido		Doente			Class. Binário	
	Indiv.	Grupo	Literal	Figurado	Autor	Outra	NA	ILA	não-ILA
Psoríase	9533	12818	21948	403	5021	316	17014	3944	18407
Urticária	18862	3087	15555	6394	12957	251	8741	7408	14541
D. Atópica	2752	1662	4323	91	1691	169	2554	1460	2954
Vitiligo	22875	32849	54208	1516	9247	3884	42593	6788	48936
Diabetes	266409	306593	514699	58303	127407	10924	434671	86258	486744
Asma	365353	52637	390214	27776	269272	16748	131970	209570	208420

Tabela 5.11: Distribuição de tweets que mencionam cada doença entre várias classes.

%	Autor		Sentido		Doente			Class. Binário	
	Indiv.	Grupo	Literal	Figurado	Autor	Outra	NA	ILA	não-ILA
Psoríase	42.7	57.3	98.2	1.8	22.5	1.4	76.1	17.6	82.4
Urticária	85.9	14.1	70.9	29.1	59.0	1.1	39.8	33.8	66.2
D. Atópica	62.3	37.7	97.9	2.1	38.3	3.8	57.9	33.1	66.9
Vitiligo	41.1	58.9	97.3	2.7	16.6	7.0	76.4	12.2	87.8
Diabetes	46.5	53.5	89.8	10.2	22.2	1.9	75.9	15.1	84.9
Asma	87.4	12.6	93.4	6.6	64.4	4.0	31.6	50.1	49.9

Tabela 5.12: Distribuição percentual de tweets que mencionam cada doença entre várias classes.

Doença	Classificadores ASD				Class. Binário				Correspondência	
	ILA	%	Não-ILA	%	ILA	%	Não-ILA	%	#	%
Psoríase	4682	20.9	17669	79.1	3944	17.6	18407	82.4	20941	93.7
Urticária	7594	34.6	14355	65.4	7408	33.8	14541	66.2	17315	78.9
D. Atópica	1600	36.2	2814	63.8	1460	33.1	2954	66.9	4008	90.8
Vitiligo	8236	14.8	47488	85.2	6788	12.2	48936	87.8	52706	94.6
Diabetes	98951	17.3	474051	82.7	86258	15.1	486744	84.9	525427	91.7
Asma	246634	59.0	171356	41.0	209570	50.1	208420	49.9	342648	82.0

Tabela 5.13: Distribuição percentual de tweets que mencionam cada doença entre várias classes.

A partir da tabela 5.12 é possível concluir que urticária e asma são as doenças com maior proporção de tweets escritos por um autor individual. Urticária é também, proporcionalmente, a doença com maior utilização figurada. De seguida, encontra-se diabetes usada frequentemente no Brasil, com trocadilhos com o adjectivo 'doce'. Asma e urticária são as doenças com maior proporção de doentes a publicar no Twitter. Quanto se trata

de mencionar a condição de saúde de outra pessoa que não o autor, vitiligo sobressai, como esperado, devido aos muitos tweets que mencionam Michael Jackson e Winnie Harlow. Asma, seguida de urticária e dermatite atópica, com considerável diferença, são as doenças com maior proporção de tweets ILA.

5.3 Resultados da filtragem com classificador binário

São agora apresentados os resultados provenientes da filtragem com o classificador binário treinado sobre o conjunto de dados complementado com Active Learning.

5.3.1 Análise de Tópicos

#	Tópico
0	urticaria me tenho anos minha rinite chega eu dermatite_atopica sinusite nome alergia asma será leio ouço segunda comi vamos causa
1	urticaria eu minha não psoriase tenho alergia meu tô mais nervosa me estou crise to dia corpo tomar frio remédio
2	psoriase minha eu não meu voltou tava voltando vitiligo bom dia merda hoje semana agora deus estresse tenho muito mão
3	psoriase eu não minha mais dermatite_atopica meu me aguento muito tô pele URLREF vida tão meus vitiligo sobre fico atacou
4	😯 psoriase dermatite_atopica couro cabeludo problemas minha feia todas contato legal descobrir dermatite pele oq fica mt acabei kkkkkkk meus
5	🤦 mentira pedir fígado 21 gostoso exames come luta idiota manter possibilidade sessão roxo campanha burro normalmente burra #mtvhottest
6	urticaria me não eu dando :(produce la meu tu mais pensar mim deus el tengo odeio causa coisas vejo
7	😂 me urticaria 😣 saiu ñ asco ha agarra dado caro lembra português tirei suas cai tinham erros sarna los
8	me vitiligo minha eu urticaria não psoriase disse 😑 mãe dermatologista dia mancha era perguntaram ligo tão outro cura vale
9	tô vitiligo mancha parece limão 8 piscina carai perna doida noticia sabiam chamou on D= cota reverso photoshop mode catapora
10	eu vitiligo tenho não minha me acho era meu psoriase URLREF sou nao agora cara pessoas mais pele mãe doença
11	vitiligo não queimadura 😪 ok pescoço velho foto manchado despelando imagens meu olhos corretivo azul hahaha perfil doar precisa efeito
12	psoriase minha tenho 😩 me viva dermatite_atopica negócio pomada dias carne mais parei dms 😱 inverno preciso leva anda acabando
13	vitiligo eu parece tenho to meu descascando minha parecendo rosto estou cara URLREF pele mano agora sol mais corpo tão
14	😭 eu vitiligo urticaria tenho menina 2 3 sou agora 1 jesus chama bunda pareço mundo so diz falam pro
15	😇😞 urticaria 😩 música 😠 beijo tl chá faltaba atinge caras maria lo musica acne estao tomou bolo tmb
16	😢😣 tenham máximo quente ❤️ banho kkkkkkkkkkkkkk tenho 😕 ✌️ gelado conhecida presencia psoriase precisam urticaria irei meu
17	psoriase você minha vida sua cu URLREF causas sintomas 💔 auto tratamentos conhecer leve estima motivo fodendo cura boas melhor
18	dermatite_atopica não psoriase pior odeio ninguém desejo minha vida porra eh saco cura pele doença horrível nao ruim merece :/
19	vitiligo meu minha ela pai eu coça contrário falou negra psoriase celular eles mãe filha diagnóstico igual tv tratamento

Tabela 5.14: 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelo **classificador binário** treinado com os dados de **crowdsourcing e active learning**.

Procede-se agora à análise de tópicos listados na tabela 5.14.

O tópico 0 faz referência a urticária e dermatite atópica, bem como pronomes e formas verbais correspondentes à primeira pessoa do singular (‘me’,‘minha’,‘eu’,‘leio’,‘ouço’,‘comi’). O que mais sobressai neste tópico é a referência a rinite e sinusite. No caso de a pessoa sofrer de alergias sazonais, tanto rinite, como sinusite e urticária (não crónica) se podem

fazer presentes simultaneamente. Existe também certa associação entre dermatite atópica, rinite alérgica e asma. Num dos sites do Serviço Nacional de Saúde português [3], podemos ler: *"A dermatite atópica pode estar associada a doenças como asma e rinite alérgica, sendo desencadeada pela interação de fatores genéticos e ambientais."*. O termo 'comi' também pode sugerir uma alergia a um alimento.

O tópico 1, à semelhança do tópico anterior, contém elementos que indicam a referência dos autores ao próprio estado de saúde ('eu','minha','tenho','meu','tô'/'to'/'estou', 'me'). Do termo 'nervosa' se pode assumir o estado emocional, bem como o sexo do autor.

O tópico 2 sugere tratar-se da reincidência de uma das dermatoses, psoríase ou vitiligo, ou pelo menos, do agravamento dos sintomas devido ao stress ('estresse'), confirmado pelo recurso ao calão ('merda'). A 'mão' consite possivelmente numa região do corpo frequentemente afectada.

O tópico 3 vai ao encontro dos tópicos anteriores quanto à referência ao eatado de saúde próprio. Apenas urticária não é mencionada.

O tópico 4 contém indicadores emocionais contraditórios ('😢','kkkkkkk'), bem como adjectivos ('feia','legal'(bom, fixe, bonito)) na mesma relação.

O tópico 5 não menciona qualquer dermatose. Não foi possível traçar uma análise especulativa sobre o assunto que engloba os termos presentes, talvez se tratando de ruído.

O tópico 6 reflecte uma emoção negativa (':(') associada ao uso da palavra urticária, mas não fica evidente se o faz no sentido literal ou figurado. Palavras de língua espanhola estão presentes no tópico.

O tópico 7 tem o seguinte tweet em consideração:

> erros de português, pra mim, são imperdoáveis.. não q eu esteja imune, masss... me dá urticária!!!!s

Torna-se assim evidente que há tweets que não referem urticária no sentido literal, a serem classificados incorrectamente, não podendo assim ser filtrados.

O tópico 8 não parece sugerir um assunto em específico.

O tópico 9 não aparenta tratar do estado de saúde pessoal do autor, devido aos termos que remetem para outros assuntos ('notícia','photoshop').

O tópico 10 é bastante disperso. Inclui a palavra 'mãe', também incluída no tópico 19 desta tabela, e em todas as tabelas analisadas anteriormente. Os primeiros quatro termos parecem apontar para a negação do autor quanto à possibilidade de sofrer vitiligo.

No tópico 11, está contida a palavra 'queimadura'. Ela está presente no corpus de tweets relativo a vitiligo, 29 vezes na forma singular e 36 vezes no plural. Seguem dois dos tweets que estão possivelmente associados a este tópico:

> se não sabem, uma queimadura em uma região sensível como a que foi pode piorar muito o vitiligo, e trazer mts transtornos...

> nossa cara, minha mão tá tão feia que quem olha pensa que eu estou com vitiligo.. mas é queimadura!

[3] http://hff.min-saude.pt/dermatite-atopica-e-mais-do-que-pele-seca/

Estes tweets foram classificados como 'ILA', mas o segundo trata-se de um equívoco.

O tópico 12, entre outros termos, contém 'psoríase' e 'inverno'. Numa inspeção aos resultados da classificação surge um tweet com ambos os termos:

```
tá falando sério que o inverno só começa hoje?  aí minha psoríase
```

A palavra 'negócio' pode ser utilizada como simples sinónimo da palavra 'situação' ou 'problema', como atesta o seguinte tweet:

```
amanhã tenho consulta com a dermatologista pra ver sobre minha
psoríase.  o negócio não vai embora :/.
```

O tópico 13 é traduz os tweets onde o autor afirma que as manchas da pele decorrentes de queimaduras solares se parecem com vitiligo.

O tópico 14 aparenta conter bastante diversidade de termos. Procurando por 'jesus' no conjunto de tweets que mencionam vitiligo, foi encontrado o seguinte tweet, erradamente classificado como 'ILA':

```
há cada dia tenho motivos para ter mais e mais fé.  eu vi um
milagre...  eu vi jesus curar um garotinho que tinha vitiligo..
```

Da mesma forma, para dermatite atópica:

```
jesus amado, não desejo dermatite atópica nem para meus piores
inimigos
```

O tópico 15 inclui 'urticária' e 'música'. Uma inspecção permitiu encontrar o seguinte tweet classificado como 'ILA', erradamente, já que se trata do recurso a urticária no sentido figurado: *"é só eu escutar "shawty's like a melody in my head" que já me da urticária. eu nao suporto mais essa música. grrrr"*.

O tópico 16 inclui uma mistura de emojis de conotação positiva e negativa. Menciona as doenças de psoríase e urticária, e baseia-se, provavalemente, em recomendações ou advertências quanto a banhos quentes ou gelados.

O tópico 17 apresenta-se com evidente conotação emocional negativa ('cu','💔','fodendo'). A par com os termos '`URLREF`', 'causas', 'sintomas' e 'auto' 'tratamentos', este tópico pode ser constituído de tweets que tecem críticas a páginas populares, não especializadas, que recomendam certas práticas para tratar as doenças, podendo sujeitar os leitores a perigos ou agravamentos.

O tópico 18 parece traduzir, de forma bastante evidente, os tweets semelhantes ao último citado no tópico 14: o desejo altruista de não desejar dermatite atópica ou psoríase a 'ninguém', nunca na 'vida', por ser demasiado 'horrível'/'ruim'/ um 'saco', o que ninguém 'merece'.

O tópico 19 parece reunir alguns tweets que abordam a doença vitiligo nos familiares ('pai','mãe','filha').

Apreciação

A tabela 5.14 contém **4** ocorrências do termo '`URLREF`'. Inclui **22** elementos gráficos: 19 emojis (15 negativos e 4 positivos) e 3 emoticons (negativos). Pela análise especulativa, consideram-se ao todo 11 tópicos compatíveis com tweets ILA (0, 1, 2, 3, 4, 6, 8, 12, 16, 17 e 18).

5.3.2 Emoções

Enquanto nos gráficos do capítulo 4, as emoções positivas se mostravam presentes nos subconjuntos 'Não-ILA', e consequentemente em 'Todos', nos gráficos da figura 5.1 parece haver uma maior erradicação dessas emoções para psoríase, dermatite atópica e vitiligo. No entanto, constitui uma novidade a presença, pequena mas visível, da emoção positiva de alta intensidade (V+A+, verde).

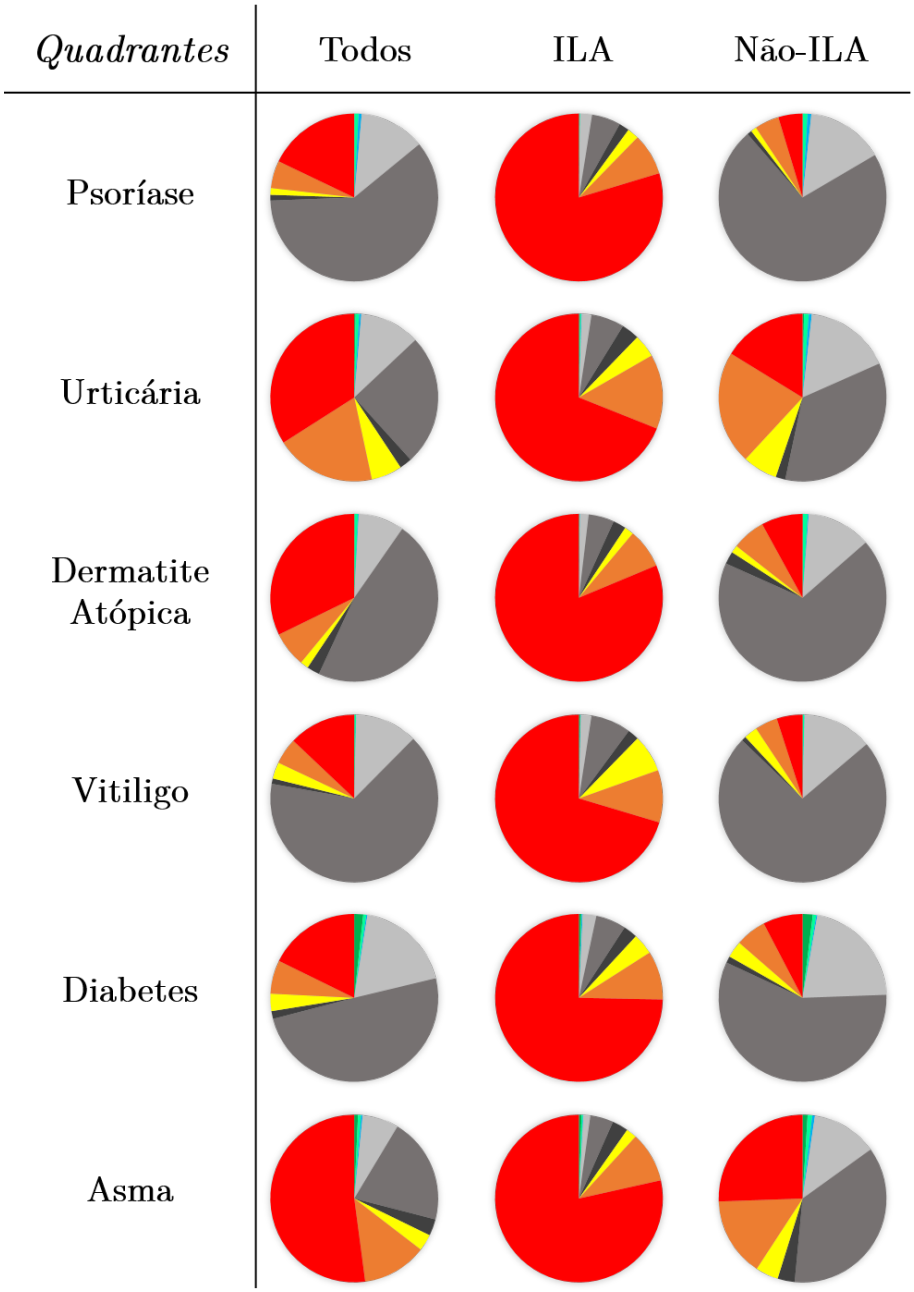

Figura 5.1: Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino recorrendo **Active Learning** e filtragem realizada tendo em conta o **classificador binário (ILA vs não-ILA)**.

As diferenças dos gráficos ILA entre as doenças, são agora mais difíceis de traçar, tendo todas aumentado a proporção representada a vermelho.

Segue-se a ordenação decrescente das percentagens dos grupos emocionais de valência negativa dos subconjuntos ILA:

V-A+ (vermelho): dermatite atópica (81.3%), psoríase (79.7%), asma (78.5%), diabetes (74.7%), vitiligo (70.4%), urticária (69.0%).

V-A0 (laranja): urticária (14.4%), vitiligo (10.1%), asma (9.7%), diabetes (9.4%), psoríase (8.0%), dermatite atópica (7.6%),

V-A- (amarelo): vitiligo (7.2%), urticária (4.4%), diabetes (4.1%), psoríase (2.2%), asma (2.0%), dermatite atópica (1.8%).

5.3.3 Distribuição dos tweets no tempo

Tecem-se agora as considerações mais relevantes, comparativamente aos resultados do capítulo anterior.

Na figura 5.2, os picos associados a Outubro de 2008 têm muito menor amplitude, comparativamente às distribuiçõs anteriores relativas a psoríase.

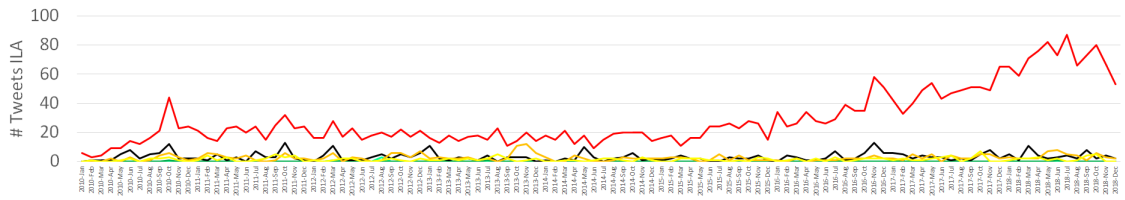

Figura 5.2: Comparação da distribuição de tweets que mencionam **psoríase**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **Active learning**.

As distribuições ao longo do tempo por emoção a partir do subconjunto ILA relativo a vitiligo em ambos os casos descritos no capítulo 4, apresentam um pico de valência negativa em Março de 2015. Foi considerado um equívoco, já que do tweet não se pode traduzir uma emoção negativa. Na figura 5.3, o pico consta na curva a preto, ou seja, foi classificado como tendo valência neutra, mostrando uma melhoria neste ponto, comparativamente ao classificador de valência do capítulo 4.

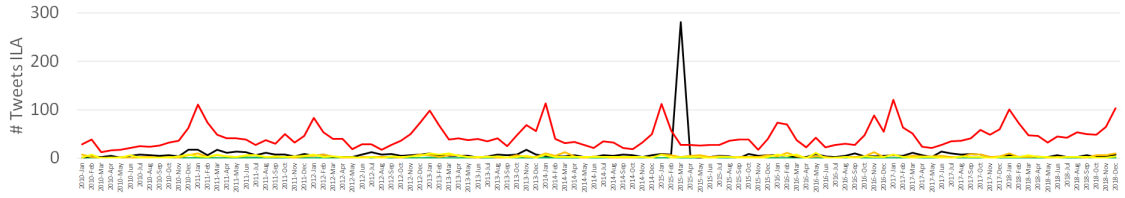

Figura 5.3: Comparação da distribuição de tweets que mencionam **vitiligo**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino complementado com **Active learning**.

Nas figuras 13 e 14, relativas a diabetes e asma, respetivamente, destaca-se a insignificâncias das restantes curvas face à curva vermelha (V-A+).

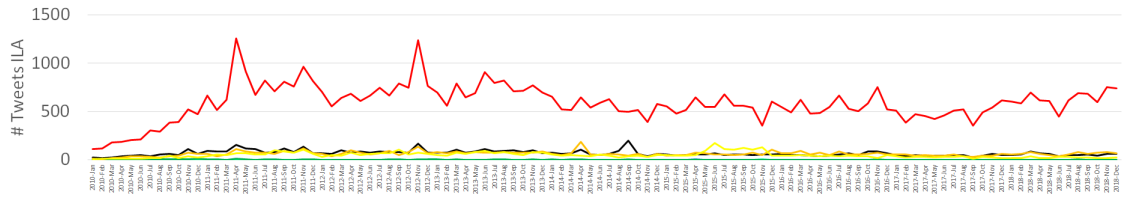

Figura 5.4: Comparação da distribuição de tweets que mencionam **diabetes**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **Active learning**.

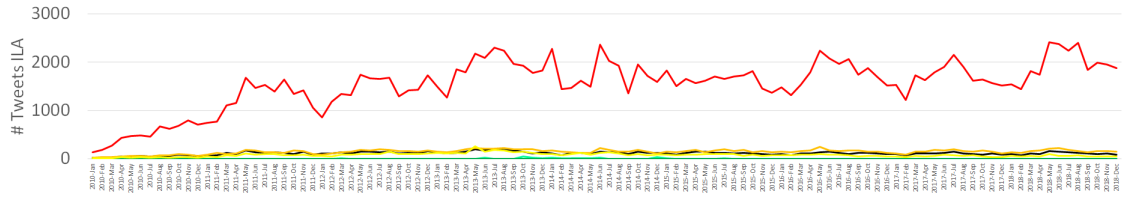

Figura 5.5: Comparação da distribuição de tweets que mencionam **asma**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **Active learning**.

5.4 Resultados da filtragem com três classificadores

Apresenta-se agora o último conjunto de resultados deste trabalho. Os gráficos da distribuição dos tweets no tempo não traduzem novidade, por essa razão, foram movidos para o anexo e este capítulo termina com a análise da distribuição global de emoções, seguindo-se a discussão final de todo o trabalho.

5.4.1 Análise de Tópicos

A partir da tabela 5.15, fazem-se os esforços para formalizar o conhecimento latente.

O tópico 0 traduz, de forma evidente, os mesmos tweets anteriormente relatados onde o autor compara as suas queimaduras solares com manchas de vitiligo.

O tópico 1 reúne todas as quatro dermatoses e expressões emocionais comuns de insatisfação, como o emoji triste, 'morrendo', 'horrível' ou 'mal'. Vitiligo é a doença que mais goza de aceitação, não constituindo sequer uma barreira na indústria da moda, como o prova a modelo Winnie Harlow, congratulada nas redes sociais pela sua beleza. O emoji apaixonado deve estar associado a vitiligo por essa razão.

O tópico 2 faz referência a urticária aquagénica. Menciona também quatro condições, distintas das dermatoses, que foram contabilizadas dentro do corpus total correspondente a cada dermatose estudada (é importante ter em conta o número de tweets que cada corpus inclui): enxaqueca (urticária: 58, psoríase: 28, dermatite atópica: 4); gripe (urticária: 37, psoríase: 22, dermatite atópica: 2); rinite (urticária: 290, psoríase: 28, dermatite atópica: 114); gastrite (urticária: 47, psoríase: 52, dermatite atópica: 3). Desta análise pode-se concluir que a gripe e a enxaqueca aparecem mais associadas a urticária, a rinite aparece associada em primeiro lugar à dermatite atópica (contém menos de 1/5 dos tweets do

#	Tópico
0	vitiligo eu parece tenho to meu descascando minha parecendo tô mancha cara não rosto branca sol agora estou pele limão
1	odeio eu 😂 dermatite_atopica vitiligo urticaria psoriase 😍 mais fazendo cara morrendo pele chamado sinto sobre horrível mal não me
2	urticaria 😞 moral dias dermatite_atopica aquagênica mata nascer ana unha make sensação enxaqueca gripe psoriase rinite gastrite #100factsaboutme maria aumentou
3	urticaria psoriase não tenho estou eu dermatite_atopica me usar 😕 URLREF feliz pele mais fui pés 😪 resultado #vcnobemestar tratamento
4	não urticaria eu me tenho URLREF ela psoriase minha gente dermatite_atopica mais vitiligo alergia causa pele meu alérgica doença muito
5	minha psoriase urticaria eu não dermatite_atopica me meu tô to merda mais vida ataca atacada tomar nervosa dia agora muito
6	não vitiligo eu mais psoriase dermatite_atopica doença queria tenho URLREF ele branco seria posso vida tratamento contagiosa saber pessoas única
7	eu psoriase vitiligo tenho não me era urticaria minha nao meu sou dermatite_atopica anos mais muito hoje tava descobri fui
8	urticaria 😭 :(puta tengo maldita tenho me 😣 D= nervosa estou coça palavra agora 1 2 pânico baixa bateu
9	urticaria 😴 não psoriase alergia saber dormir resolveu além merece tô agora mais minha me sempre conta kim dando deus
10	disse vitiligo urticaria lá perguntaram viada loca feia confusão morte criaram brancos medico vo pro livro tenham oq ela ha
11	mais aguento urticaria não psoriase cura odio mundo pior dermatite_atopica 😑 😤 medicina cuba quero nao parabéns fora bela continua
12	não psoriase desejo dermatite_atopica urticaria meu ninguém eu to estressada pro tão pior meus inferno temos chato 👌 estressado inimigo
13	psoriase semana urticaria oi meses todos filha :/ 2 dermatite_atopica crise URLREF há comecei beijo bem mais estou melhora sofrendo
14	vitiligo eu acho não urticaria me minha URLREF modelo tenho tão mais to gente tava estou teve ele ficou mim
15	vitiligo urticaria não psoriase dia agora você doença eu branco meu tenho vi pele lindo pessoas couro cabeludo eh bom
16	vitiligo meu psoriase minha URLREF me não pele eu sobre top história mais minhas sucesso vão muito deus define contar
17	:D 😢 saúde psoriase 😫 adivinha melhorando crl 😔 orgulho joelho acabar øministério finalmente estas digital caraca disseram dó
18	urticaria tenho meu vitiligo eu nesse me corpo URLREF alguem não ouço 👏 deus amanhã momento vejo desculpa mim programa
19	urticaria ❤️ verão 😡 :O vaca malhado macumba pelas aquilo sintoma ataque 😐 200 malhada amém inflamatória adquirir perde

Tabela 5.15: 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelos três classificadores **(Autor, Sentido, Doente)** treinados com os dados de **crowdsourcing e active learning**.

corpus de urticária) e em segundo à urticária. A gastrite aparece associada a psoríase e uticária num nível semelhante.

O tópico 3 inclui a referência a 'pés' pela primeira vez, em todos os tópicos analisados. A hashtag #vcnobemestar alude ao programa 'Bem-Estar', que aborda assuntos de saúde, na TV Globo, estação brasileira.

O tópico 4 descreve de forma elementar cada uma das quatro dermatoses como 'doença' de 'pele'. Algumas podem ter a sua 'causa' numa reacção 'alérgica'.

O tópico 5 menciona o estado emocional alterado ('nervosa'), que pode aumentar o grau dos sintomas ('mais','muito','atacada'), cuja influência negativa na 'vida' das pessoas é atestada pelo recurso a calão ('merda').

O tópico 6 parece sugerir a necessidade de se fazer 'saber' às 'pessoas' de que nenhuma destas dermatoses é 'contagiosa'.

O tópico 7 é compatível com a representação de tweets em que o autor descreve a sua condição dermatológica, dada a presença de pronomes, principalmente possessivos, e

formas verbais correspondentes à primeira pessoa do singular.

O tópico 8 é rico em recursos gráficos para a expressão de emoções, contendo dois emojis e dois emoticons, os quatro de conotação negativa. Para além disso ainda recorre ao calão ('puta'), e menciona um adjectivo ('maldita') e outros estados emocionais desagradáveis ('nervosa', 'pânico').

Ao contrário do tópico anterior, o tópico 9 sugere a possível recuperação ('resolveu') da condição de urticária, acompanhada do emoji sorriso.

O tópico 10 inclui insultos, mas não é evidente ao que se destinam.

O tópico 11 contém referência a Cuba. Uma inspecção nos corpus de cada uma das três dermatoses mencionadas (urticária, psoríase, dermatite atópica) permitiu concluir que não existe interligação entre elas. Por exemplo, no caso de urticária, existem tweets que citam a doença no sentido figurado para caracterizar o fim do programa 'Mais Médicos', pelo qual se encontravam médicos cubanos a operar no Brasil.

O tópico 12 vai ao encontro do tópico 18 da tabela 5.14 onde os autores, portadores de dermatose, expressam não a desejar ao pior inimigo.

O tópico 13 contém um conjunto de termos bastante heterogéneos, ora sugerindo recuperação ('melhora','bem') ou sofrimento ('sofrendo','crise','`:/`').

O tópico 14 parece misturar menções ao estado de saúde pessoal com referências à modelo conhecida por ter vitiligo, indiciadas pela presença de ligações externas ('`URLREF`').

O tópico 15 inclui referência ao 'couro cabeludo', provavelmente associada a vitiligo, já que também consta a palavra 'branco' que pode reflectir a aparência dos cabelos. O couro cabeludo pode ser alvo de uma grande variedade de dermatoses.

O tópico 16 sugere a presença de tweets que relatam a 'história' de 'sucesso' de quem sofre de vitiligo ou psoríase. Essa história pode se tratar da luta contra o estigma social, ou por outro lado, pela recuperação da doença através de um tratamento eficaz.

O tópico 17 começa com um emoticon sorriso ('`:D`'), possivelmente associado ao termo 'melhorando', uma vez que se tratam de termos positivos. Ao mesmo tempo, estão presentes três emojis, todos de carácter negativo, bem como a abreviação 'crl', comum do calão.

Nada em especíco parece sugerir o tópico 18.

O tópico 19 é bastante particular: menciona o 'verão', que pode ser uma altura benéfica ('❤') para quem tem urticária, ou prejudicial ('😡') para quem tem urticária. Não é um erro de digitação: tudo depende de qual tipo de urticária sofre a pessoa, colinérgica ou de frio. Não obstante tamanha contradição de emojis, ainda se evoca o padrão 'malhado' para a 'vaca' e se espera que não se trate de um insulto a quem sofre de vitiligo. Faltava ainda só misturar bruxaria ('macumba') com religião ('amém'/amen).

Apreciação

A tabela 5.15 contém **7** ocorrências do termo '`URLREF`'. Inclui 22 elementos gráficos: 18 emojis (11 negativos, 3 neutros e 4 positivos) e 5 emoticons (1 positivo, 1 neutro, 3 negativos). Entre os tópicos, admite-se compatibiliade com tweets ILA em **8**: 1, 2, 4, 5, 6, 7, 12, 16. Poderão ser considerados confusos **7** tópicos: 3, 8, 9, 13, 7, 18, 19.

5.4.2 Emoções

A mesma tendência se conserva: dermatite atópica, asma e psoríase, lideram nas proporções de tweets ILA com emoções de valência negativa e alta ativação.

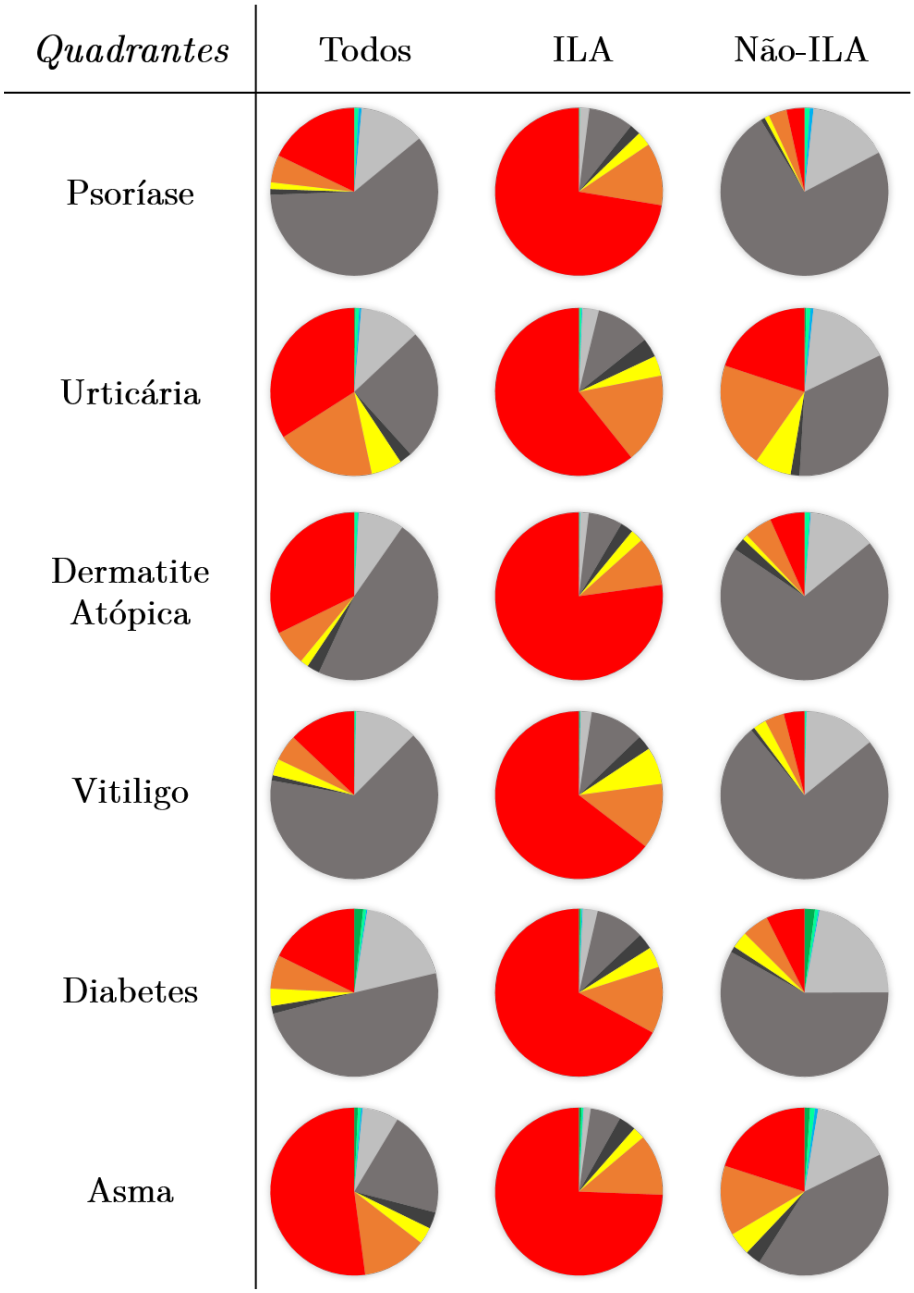

Figura 5.6: Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino recorrendo **Active Learning** e filtragem realizada tendo em conta os **três classificadores (Autor, Sentido, Doente)**.

V-A+ (vermelho): dermatite atópica (77.2%), asma (74.4%), psoríase (72.4%), diabetes (67.1%), vitiligo (64.6%), urticária (60.7%).
V-A0 (laranja): urticária (17.4%), diabetes (12.9%), vitiligo (12.6%), psoríase (12.1%), asma (11.8%), dermatite atópica (9.4%).
V-A- (amarelo): vitiligo (7.2%), diabetes (4.0%), urticária (3.9%), psoríase (2.8%), dermatite atópica / asma (2.4%).

Capítulo 6

Discussão e Conclusão

Com o intuito de perceber como as pessoas expressam as suas condições de saúde, nomeadamente as psicodermatoses, nas redes sociais, foi recolhido um conjunto de tweets de aproximadamente 1 milhão de exemplares. Foi realizada uma análise LDA sobre as psicodermatoses e se chegou à conclusão que existe uma grande presença de tweets que não são da autoria de doentes. Foram realizados alguns testes recorrendo a léxicos de emoções, para texto e emojis/emoticons, mas porque as redes sociais são caracterizadas por uma gíria própria, tornou-se necessário procurar uma alternativa, já que o léxico de texto requer palavras escritas formalmente. Os tweets foram dispostos ao longo do tempo (número de tweets por mês, desde 2010 a 2018) e verificaram-se episódios interessantes em todos as psicodermatoses abordadas. Foi ainda averiguado se era possível restringir a análise a tweets provenientes de Portugal e/ou do Brasil, no entanto o mecanismo para apurar a nacionalidade não se revelou suficiente. Pela necessidade de filtrar os tweets e de obter um sistema para reconhecer o tipo de emoção presente, recorremos a uma plataforma online de *crowdsourcing*, criando um conjunto anotado com cerca de 500 tweets. Foram admitidas duas modalidades de filtragem, uma com um classificador binário e outra a partir de três classificadores, cada um com seu problema específico. Realizou-se a filtragem com ambas as modalidades e apuraram-se os resultados para análise de tópicos e espectros emocionais, agora filtrados. Foram dispostos no tempo os subconjuntos filtrados, bem como a sua divisão pelas emoções presentes. Na procura por resultados mais consistentes, anotaram-se mais tweets com recurso a active learning. As experiências foram repetidas, e tecidas as considerações sobre alguns pormenores que evidenciavam falhas nos resultados sem recurso a active learning. Pudemos concluir uma evolução dos resultados da análise de tópicos, realizada inicialmente, para os resultados com tweets filtrados, que evidenciam uma maior relevância de tweets da autoria de doentes, que falam concretamente das suas doenças. Da análise de emoções, os resultados apontam para uma prevalência de emoções semelhantes a raiva e frustração em todos os casos, mas as doenças de dermatite atópica, asma e psoríase, apresentam os maiores níveis. Diabetes e asma apresentam-se globalmente como as doenças mais populares entre as abordadas. As maiores percentagens de tweets ILA fazem-se presentes no caso da asma, seguida por urticária e dermatite atópica. Seguem agora algumas considerações mais específicas.

Os objetos de estudo neste trabalho são publicações do Twitter, textos informais caracterizados pela presença de erros ortográficos e gramaticais. Seria possível aplicar um corretor automático durante o processamento de texto, mas sem verificação correr-se-ia o perigo de deturpar o conteúdo, especialmente pela ocorrência de palavras mais incomuns associadas às doenças e, se um verificador tomar em conta o contexto, pode ser influenciado pela presença de emojis e emoticons, não compreendendo a estrutura sintática.

Em determinado momento do estudo, foi averiguado realizar uma listagem de todos os unigramas, tentando-os agrupar por semelhança e criar um dicionário de ortografia correta, adicionando a correção como um passo no pré-processamento. A dificuldade seria a duração do processo, porque as palavras fora do contexto obrigam à inspeção do conjunto de dados para perceber qual a tradução correta a fazer. O número de features ficaria certamente reduzido, mas alguns termos errados não poderiam ser contemplados, por serem compatíveis com várias palavras diferentes.

Outra etapa de processamento de texto poderia ter sido realizada. À semelhança do que foi feito com emojis e emoticons, convertidos aos seus aliases e anexados a um identificador do elemento gráfico, poderia ter sido feito o mesmo em relação à pontuação, especialmente pontos de exclamação, que guardam informação emocional, de forma a preservá-lo durante o processo de tokenização. A repetição de letras também poderia ser normalizada. Aqui o perigo de deturpação seria menor do que a perda total de informação que acabou por ocorrer, por exemplo, no caso da gargalhada ('kkkkk','hahaha'), expressa por um conjunto sucessivo de letras ou pares, muito diverso em número. Agrupar estas ocorrências numa só feature teria sido mais apropriado. Quanto à contabilização das letras maiúsculas em palavras (não em emoticons), que podem também guardar informação emocional, o processamento para as considerar seria mais difícil, porque a conversão a letra minúscula é elementar para gerar generalização, aplicabilidade. Talvez existam métodos para contabilizar as letras maiúsculas presentes. Uma possibilidade seria duplicar a presença da palavra que usa letra maiúscula.

Começando por abordar a representatividade do conjunto anotado, para o construir poderia ter sido realizada uma seleção aleatória sobre o conjunto de dados total. No processo de active Learning foi adoptado o algoritmo `MultinomialNB` como classificador e a incerteza da classificação como critério de selecção do tweet para anotação, sendo devolvido aquele com maior incerteza. No início foram feitas algumas experiências com estes parâmetros, não sendo o total dos tweets anotados decorrentes deste par. De facto, a forma como esta estratégia de anotação foi implemantada aqui não é recomendável, talvez logo de início deveriam ser consideradas as 6 doenças. O processo de active learning foi dividido para cinco problemas (autor, sentido, doente, valência, ativação) e cerca de 300 tweets foram anotados. Logo, apenas foram devolvidos perto de 60 tweets por cada problema de classificação, tendo em conta o contributo que lhe dariam. Considerando que o processo não esteve corretamente implementado desde o início, e por isso, uma considerável quantidade de tweets não corresponde aos sugeridos para anotação, o contributo adicionado ao conjunto de dados obtido de crowdsourcing é relativamente reduzido.

Para aprimorar a classificação, teria sido apropriada uma seleção de features. No entanto, com um conjunto anotado relativamente pequeno, esta etapa não foi tida em conta, embora se tenham feito experiências de relevância de features a partir da técnica estatística do χ^2, da qual se concluiu a relevância da contabilização do termo '`URLREF`' para a maioria dos problemas de classificação. O número de features usado por cada classificador dependeu do tipo de features selecionado: se lemas, o número seria menor; se não fossem contemplados bigramas, o número seria menor; da mesma forma se as stopwords[1] fossem excluídas ou os acentos eliminados. Para além disto, as features que não constam pelo menos duas vezes no conjunto de treino, também não são tidas em consideração.

Tendo em conta o nível de correspondência de resultados das modalidades de filtragem de tweets ILA, e o nível de informação que cada um pode agregar, fica a possibilidade de que o conjunto de treino de cada problema talvez beneficiasse incluir as *labels* dos restantes

[1] `https://github.com/stopwords-iso/stopwords-pt/blob/master/stopwords-pt.txt`

problemas como *features*, por exemplo, o classificador de *Doente* poderia beneficiar se soubesse à partida que um dado tweet incluído no conjunto de treino evoca a doença num sentido figurado. Esta abordagem não ficaria livre de desvantagens, já que um autêntico doente pode evocar a sua doença também no sentido figurado. De qualquer forma, só a avaliação de tal treino poderia esclarecer qual a melhor abordagem.

6.1 Sobre os Resultados

No que toca à filtragem de tweets ILA, a utilização do classificador binário apresentou melhores indícios de sucesso: o número de tópicos que incluem o termo `URLREF` foi menor para ambos os conjuntos de treino, relativamente à filtragem com três classificadores; o número de elementos gráficos (emojis e emoticons) presentes na globalidade dos tópicos foi superior, com o conjunto de treino exclusivo a crowdsourcing, tendo obtido 22 elementos, mais 4 do que a partir da filtragem ASD. Tanto no capítulo 4 como no 5, o classificador binário conseguiu discernir o episódio de publicidade abusiva, presente no conjunto relativo a vitiligo. Quanto ao número de tópicos admitidos como compatíveis com tweets ILA, as conclusões são difíceis de tecer, pela presença de tópicos confusos, cuja validade para a contabilização é muito discutível. Se desprezados os tópicos identificados como confusos, seguiria na mesma direção, apontando o classificador binário como mais adequado. Ter em conta as medidas de desempenho obtidas para os classificadores que compõem as duas modalidades, não se revela possível para apurar o melhor, porque não estão a ser retiradas da mesma aplicação. O ideal teria sido agrupar os três classificadores (Autor, Sentido e Doente) num só objeto, e o sujeitar à mesma validação do classificador binário. Nesse caso, admitir as medidas de desempenho para comparar ambas as modalidades seria o apropriado.

Analisando os resultados da tabela 6.1, podemos concluir que a menor proporção de tweets ILA se encontra relacionada à doença de vitiligo. As percentagens de tweets ILA obtidas para asma superam sempre metade dos resultados. Restringindo ao grupo das dermatoses, essa marca aplica-se à urticária. Tendo em conta as diversas percentagens apresentadas na mesma tabela, podemos concluir que os resultados obtidos da filtragem através de três classificadores treinados no conjunto exclusivo a crowdsourcing, são bastante diferentes tendo em conta as outras três modalidades que têm resultados semelhantes entre si.

De acordo com Svensson et al. (2018), a prevalência das dermatoses estudadas na Europa seria ordenada da seguinte forma: urticária, dermatite atópica, psoríase, vitiligo. Agora focando nos valores absolutos, as quatro modalidades apresentam dermatite atópica em último lugar, pela razão de não incluir todos os tweets que se referem à doença. Também apresentam a urticária com maior popularidade (ILA) do que a psoríase, indo ao encontro do estudo citado, neste ponto. O resultado com recurso a active learning e classificador binário (AL Bin.) é o único a apresentar vitiligo com menor popularidade que urticária.

Quanto à analise de tópicos, muitos deles são de difícil compreensão. A técnica LDA exige que seja predefinido o número de tópicos. Os resultados apresentados contêm 20. Podiam ser otimizados se fossem utilizados algoritmos para estimar o número ideal, mas continuaria a ser preferível manter o número de tópicos igual para as 5 análises realizadas.

No que toca aos classificadores de emoção, pode ser tecida uma consideração sobre o classificador de valência descrito no capítulo 5 (AL): classificou corretamente o pico de 270 tweets em Março de 2015 como neutro, presente no conjunto relativo a vitiligo;

Popularidade ILA	**CS Bin.**	**%**	**CS ASD**	**%**	**AL Bin.**	**%**	**AL ASD**	**%**
Psoríase	4.769	21.3	5.898	26.4	3.944	17.6	4.682	20.9
Urticária	8.059	36.7	14.726	67.1	7.408	33.8	7.594	34.6
D. Atópica	1.613	36.5	1.940	44.0	1.460	33.1	1.600	36.2
Vitiligo	9.780	17.6	15.044	27.0	6.788	12.2	8.236	14.8
Diabetes	120.152	21.0	173.946	30.4	86.258	15.1	98.951	17.3
Asma	232.003	55.5	307.195	73.5	209.570	50.1	246.634	59.0

Tabela 6.1: Popularidade de cada doença, relativa aos subconjuntos ILA, filtrados por classificador Binário (Bin.) e três classificadores (ASD), treinados sobre o conjunto exclusivo a crowdsourcing (CS) ou complementado com active learning (AL).

enquanto que, no treino exclusivo com tweets anotados em crowdsourcing (capítulo 4), os tweets deste episódio foram classificados com valência negativa (V-) e média ativação (A0), representados a laranja. Se estes indícios fossem suficientes para concretizar uma escolha, seria selecionada a modalidade que usa o conjunto complementado com active learning para o treino dos classificadores de emoção e do classificador binário. No entanto, estas observações tratam de eventos particularmente percetíveis, fáceis de identificar. A verdadeira dificuldade está em identificar os equívocos camuflados, que não evidenciam a sua presença.

Das quatro modalidades aplicadas ao longo dos capítulos 4 e 5, um resultado se apresenta comum em todas: o conjunto de tweets ILA relativo a dermatite atópica é aquele com maior proporção de tweets classificados com V-A+ (valência negativa e ativação alta), e é sempre seguido pelos conjuntos relativos a psoríase e asma. Surpreendentemente, as proporções mais significativas do mesmo quadrante no caso dos léxicos, correspondem às mesmas doenças. O conjunto obtido para a dermatite atópica é muito limitado, porque as pessoas normalmente só se referem à doença pelo termo 'dermatite' ou 'eczema' que abarcam vários tipos. Desta maneira é difícil assumir que o conjunto obtido seja realmente representativo. Os conjuntos que mencionam asma e psoríase têm tamanhos consideráveis, principalmente asma, pelo que as tendências observadas devem ser tidas em conta.

Na tabela 6.2, podemos comprovar: o número de tópicos que mencionam '`URLREF` diminuiu com a classificação; o número de emojis e emoticons, ausentes em publicações de grupos e organizações, aumentou; bem como o número de tópicos compatíveis com tweets ILA.

O número de tópicos compatíveis com tweets ILA constitui, evidentemente, uma grandeza quantitativa. Mas o facto de pouco ultrapassarem metade do número total de tópicos não indica que a tarefa foi desempenhada apenas em 50%, isto porque, a análise de tópicos pretende mostrar os assuntos que estão presentes no conjunto. Uma grande quantidade de tweets pode estar associado ao mesmo tópico. A diferença que se mostra aqui é de carácter qualitativo, mostrando que houve tópicos a perder relevância (Não-ILA), para outros serem incluídos (ILA).

Para concluir, assumindo como resultados mais fidedignos aqueles decorrentes do treino complementado com active learning e filtragem a partir do classificador binário, os resultados sugerem a condição de dermatite atópica como a mais associada a emoções negativas de alta intensidade (com todas as reticências apontadas anteriormente), seguida de psoríase, que se faz acompanhar de diabetes e asma nas mesmas proporções, e por fim, vitiligo e urticária. Destacam-se os resultados para a doença de diabetes: como doença adoptada para controlo, poderia não ser óbvio tamanha proporção associada a emoções negativas. O Professor Doutor Pedro Augusto de Melo Lopes Ferreira, da Faculdade de Economia da Universidade de Coimbra, aponta este resultado para o facto de diabetes ser

Modalidade	# URLREF	# emojis/emoticons	# Tópicos ILA	Publicidade Abusiva	Emoção Março 2015
Originais	9	6	1 (1)	-	-
CS Bin.	5	22	8	filtrado	V-A0
CS ASD	6	18	7 (3)	não filtrado	V-A0
AL Bin.	4	22	11	filtrado	neutro
AL ASD	7	22	8 (7)	filtrado	neutro

Tabela 6.2: Comparação dos resultados das quatro modalidades de classificação estudadas nos capítulos 4 (CS: crowdsourcing) e 5 (AL: active learning), com os resultados originais do capítulo 3. '#' designa 'número'. Entre parênteses está incluído o número de tópicos confusos, eventualmente compatíveis com tweets ILA. Dois agrupamentos de colunas: resultados observados pela análise de tópicos e eventos detetados na distribuição temporal relativa a vitiligo.

uma doença de presença geralmente silenciosa, ao contrário das dermatoses que se fazem notar naquele que é o órgão mais exposto do organismo humano - a pele, mas que, quando atinge patamares de grande gravidade, caracterizados por feridas, úlceras, de muito difícil cicatrização, o contributo negativo na vida dos doentes é caracterizado por um sofrimento intenso. Enquanto as dermatoses, mesmo com menor gravidade, já traduzem um estigma social que afecta a vida dos doentes, a diabetes tende a fazê-lo apenas sob considerável gravidade. Esses episódios devem ser os responsáveis pela partilha na rede social. Dado que diabetes conta com um conjunto de dados muito considerável, 136 tweets anotados podem não ser suficientes para consolidar os resultados relativos à doença, já que foram identificadas bastantes referências à doença no sentido figurado que podem não ser discernidas pelos classificadores. Quanto à doença de asma, é caracterizada por episódios aflitivos de falta de ar, sendo a associação com emoções negativas evidente. Por fim, é possível concluir que a rede social Twitter é consideravelmente utilizada para campanhas de informação e sensibilização, principalmente nos dias comemorativos de psoríase e diabetes, caracterizados pelos picos mais evidentes dos gráficos de distribuição temporal, sem filtragem.

6.2 Trabalho Futuro

O desempenho real de um classificador tenderá a ser melhor se o conjunto de treino constituir uma amostra consistente do universo de objetos que serão alvos da classificação. A abordagem com active learning permite a cada iteração, avaliar o conjunto e devolver a instância que, segundo um determinado critério, constitui a classificação mais incerta. Era de interesse apurar se existem formas de selecionar um conjunto para anotação de uma forma mais estruturada, não aleatória. Fica o desejo de averiguar se a realização de uma tarefa de clustering sobre um conjunto de tweets definida com, por exemplo, um número total de 500 clusters à priori, tornará possível a obtenção de um conjunto de 500 tweets para anotação, cada um retirado de um cluster, com grande representatividade.

No que toca aos algoritmos de classificação, na atualidade, o recurso a abordagens com redes neuronais artificiais, tem ganhado robustez nos últimos anos. Seria interessante testar sobre os conjuntos apurados.

Sobre a obtenção de resultados, seria prudente agregar as distribuições de vários anos num único só, para evidenciar eventuais ocorrências em certas épocas do ano. Dentro dos conjuntos ILA, pode ainda ser feita uma pesquisa de referências a sintomas e medicamentos, tendo também em conta a emoção apurada nos tweets que os citam, com o objetivo de avaliar se a informação presente pode ser útil para hospitais ou laboratórios, constituindo

uma fonte de feedback sobre as soluções produzidas.

O potencial das anotações recolhidas neste trabalho não foi extinto. O classificador do tipo de autor também identifica os tweets da autoria de grupos ou organizações. Com base nisso, pode ser feito um estudo para estimar a necessidade de criar uma plataforma de verificação online, à semelhança dos atuais verificadores de *fake news*, para assinalar as páginas que, sem conhecimento médico-farmacêutico especializado, aconselham tratamentos perigosos para certas doenças, motivando uma automedicação prejudicial. o objetivo seria criar um índice de confiabilidade científica.

A análise emocional realizada neste trabalho pode também ser obtida por utilizador, isolando os tweets da sua autoria que mencionam doenças, de forma a compreender a prevalência de uma determinada emoção ao longo do tempo. Deste modo, a associação entre as emoções presentes e a caraterização do seu estado emocional generalizado, seria mais consistente. No final, o espectro emocional de cada doença pode ser obtido contabilizando a emoção mais frequente por utilizador, podendo-se restringir àqueles cujo número de tweets ultrapassa um mínimo estipulado.

Bibliografia

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beguerisse-Díaz, M., McLennan, A. K., Garduño-Hernández, G., Barahona, M., and Ulijaszek, S. J. (2017). The 'who' and 'what' of #diabetes on Twitter. *DIGITAL HEALTH*, 3:2055207616688841. PMID: 29942579.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bobicev, V. and Sokolova, M. (2018). Thumbs Up and Down: Sentiment Analysis of Medical Online Forums. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 22–26, Brussels, Belgium. Association for Computational Linguistics.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Citeseer.

Criado, R., Criado, P., Sittart, J. d. S., Pires, M., de Mello, J., and Aun, W. (1999). Urticária e doenças sistêmicas. *Revista da Associação Médica Brasileira*, 45(4):349–356.

Duncan, K. O. and Koo, J. Y. M. (2018). *Psychocutaneous Diseases*, volume 1, chapter 7, pages 128–137. Elsevier, 4th edition.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Ezequiel, O. d. S., Gazeta, G. S., and Freire, N. M. A. d. S. (2007). Prevalência dos atendimentos por crises de asma nos serviços públicos do Município de Juiz de Fora (MG). *Jornal Brasileiro de Pneumologia*, 33:20 – 27.

Flekova, L., Lampos, V., and Cox, I. (2018). Changes in psycholinguistic attributes of social media users before, during, and after self-reported influenza symptoms. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 17–21.

Gohil, S., Vuik, S., and Darzi, A. (2018). Sentiment analysis of health care tweets: review of the methods used. *JMIR public health and surveillance*, 4(2):e43.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Lewis, D. D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Malheiro, R., Gonçalo Oliveira, H., Gomes, P., and Paiva, R. P. (2016). Keyword-Based Approach for Lyrics Emotion Variation Detection. In *8th International Conference on Knowledge Discovery and Information Retrieval*, KDIR'2016.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K., and Gonzalez, G. (2014). Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014:924–933.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Rodrigues, D., Prada, M., Gaspar, R., Garrido, M. V., and Lopes, D. (2018a). Lisbon Emoji and Emoticon Database (LEED): Norms for emoji and emoticons in seven evaluative dimensions. *Behavior research methods*, 50(1):392–405.

Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2014). LemPORT: a High-Accuracy Cross-Platform Lemmatizer for Portuguese. In Pereira, M. J. V., Leal, J. P., and Simões, A., editors, *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 267–274, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2018b). NLPPort: A Pipeline for Portuguese NLP. In *Proceedings of 7th Symposium on Languages, Applications and Technologies (SLATE 2018)*, volume 62 of *OASIcs*, pages 18:1–18:9, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Russell, J. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Santos, J. C. and Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling*, 11(1):S6.

Scanfeld, D., Scanfeld, V., and Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3):182–188.

Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Sinnenberg, L., DiSilvestro, C. L., Mancheno, C., Dailey, K., Tufts, C., Buttenheim, A. M., Barg, F., Ungar, L., Schwartz, H., Brown, D., Asch, D. A., and Merchant, R. M. (2016). Twitter as a Potential Data Source for Cardiovascular Disease Research. *JAMA Cardiology*, 1(9):1032–1036.

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., and Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, 44(1):256—-269.

Svensson, A., Ofenloch, R., Bruze, M., Naldi, L., Cazzaniga, S., Elsner, P., Goncalo, M., Schuttelaar, M.-L., and Diepgen, T. (2018). Prevalence of skin disease in a population-based sample of adults from five European countries. *British Journal of Dermatology*, 178(5):1111–1118.

Wood, I. and Ruder, S. (2016). Emoji as emotion tags for tweets. In *Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia*, pages 76–79.

Yin, Z., Fabbri, D., Rosenbloom, S. T., and Malin, B. (2015). A Scalable Framework to Detect Personal Health Mentions on Twitter. *J Med Internet Res*, 17(6):e138.

Anexos

Anexos - Capítulo 3

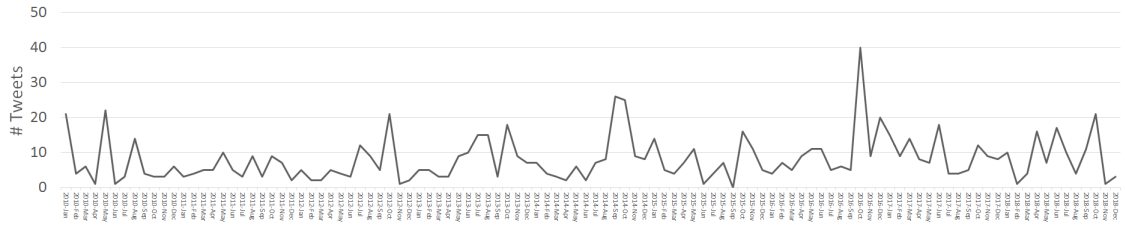

Figura 1: Distribuição temporal de tweets referentes a **Psoríase** entre 2010 e 2018, localizados em **Portugal**.

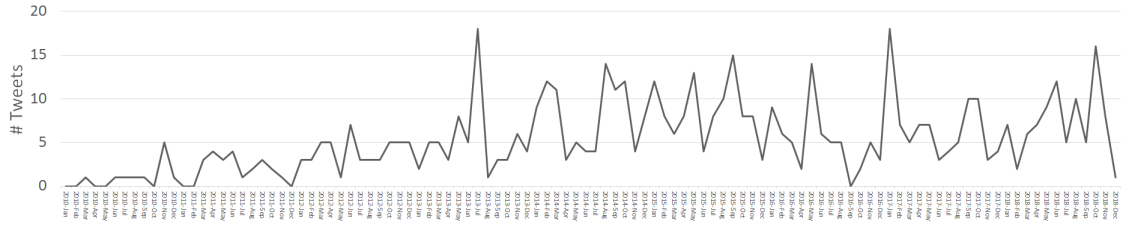

Figura 2: Distribuição temporal de tweets referentes a **Urticária** entre 2010 e 2018, localizados em **Portugal**.

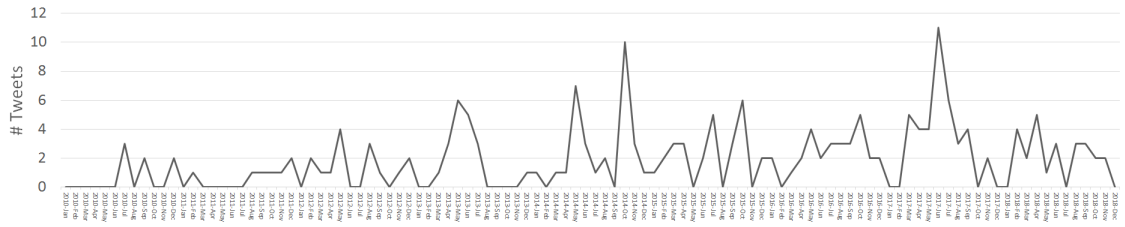

Figura 3: Distribuição temporal de tweets referentes a **Vitiligo** entre 2010 e 2018, localizados em **Portugal**.

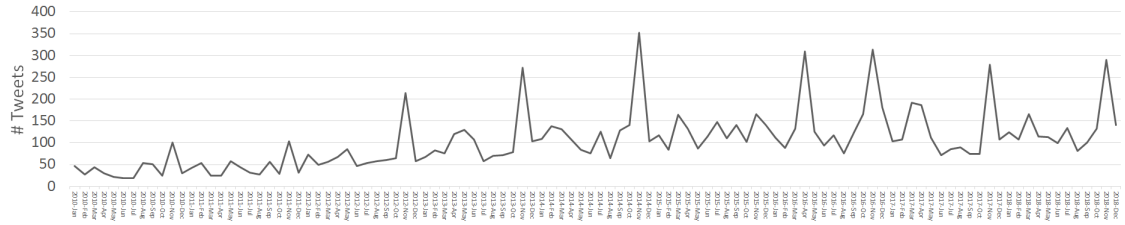

Figura 4: Distribuição temporal de tweets referentes a **Diabetes** entre 2010 e 2018, localizados em **Portugal**.

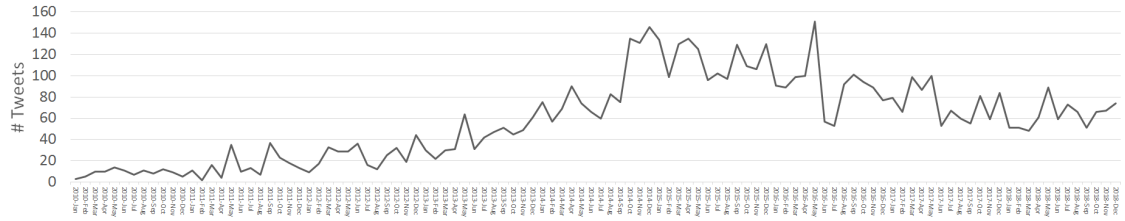

Figura 5: Distribuição temporal de tweets referentes a **Asma** entre 2010 e 2018, localizados em **Portugal**.

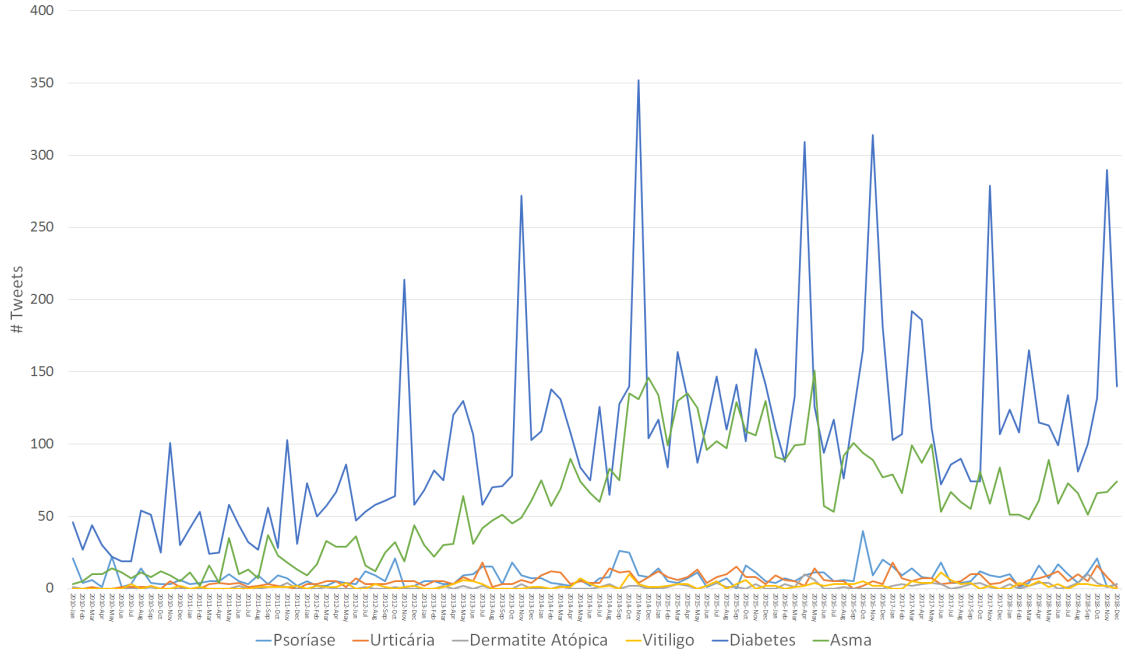

Figura 6: Comparação da distribuição temporal de tweets referentes às seis doenças abordadas entre 2010 e 2018, localizados em **Portugal**.

Anexos - Capítulo 5

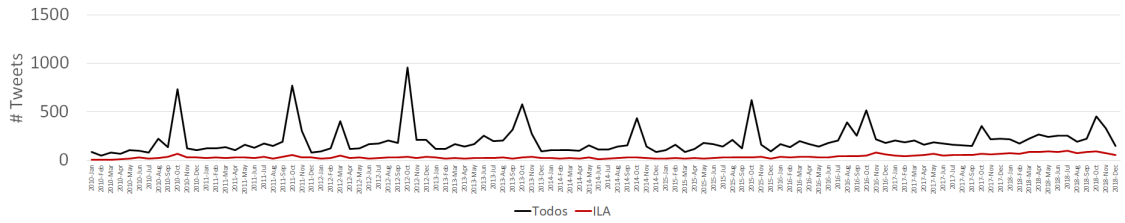

Figura 7: Comparação da distribuição de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**.

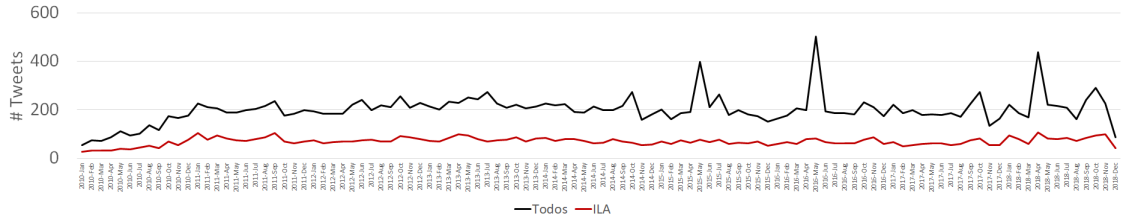

Figura 8: Comparação da distribuição de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**.

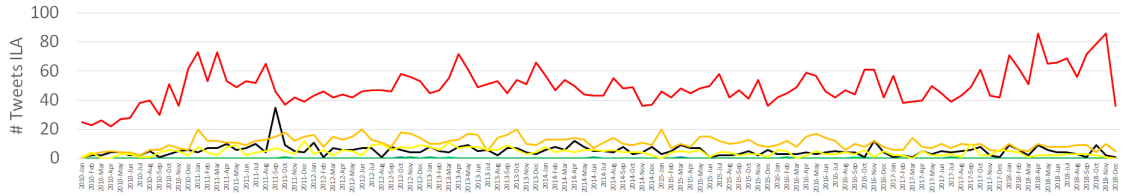

Figura 9: Comparação da distribuição de tweets que mencionam **urticaria**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

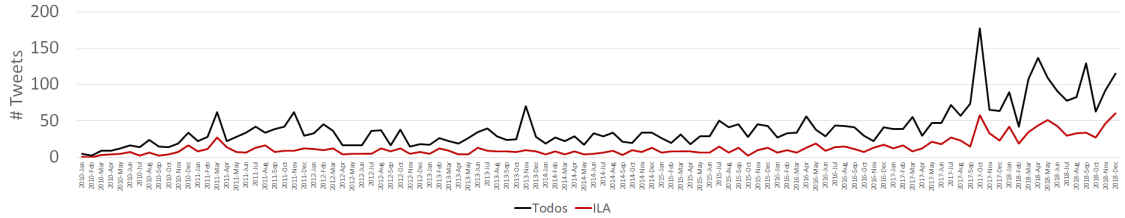

Figura 10: Comparação da distribuição de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**.

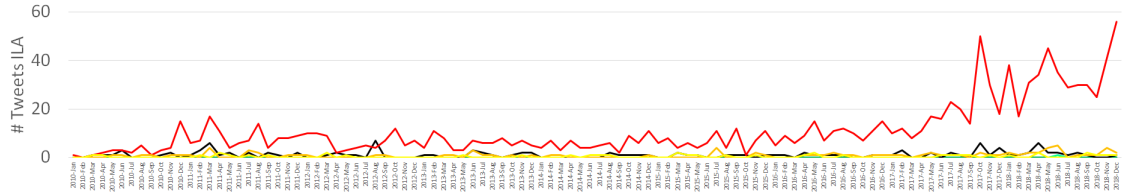

Figura 11: Comparação da distribuição de tweets que mencionam **dermatite atópica**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

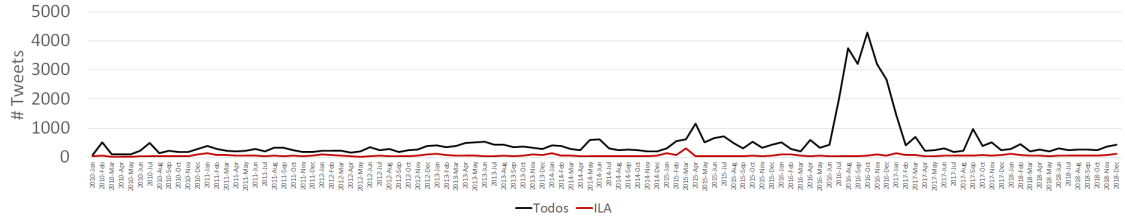

Figura 12: Comparação da distribuição de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**.

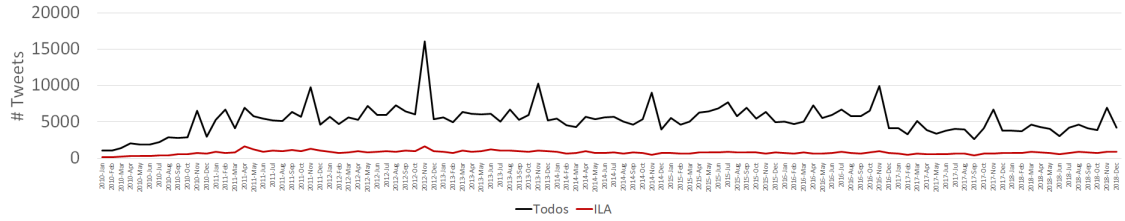

Figura 13: Comparação da distribuição de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**.

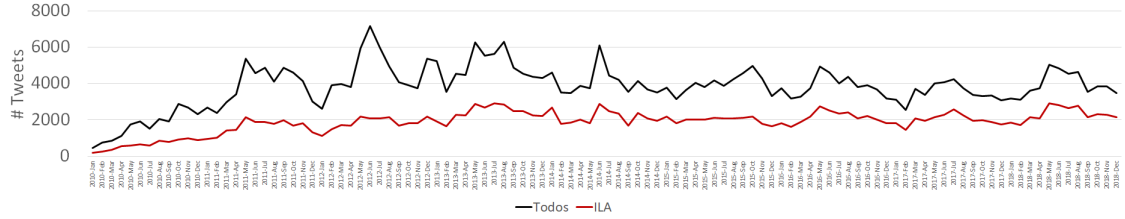

Figura 14: Comparação da distribuição de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**.

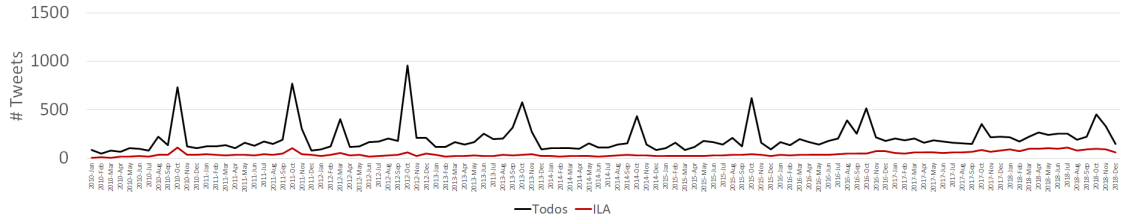

Figura 15: Comparação da distribuição de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**.

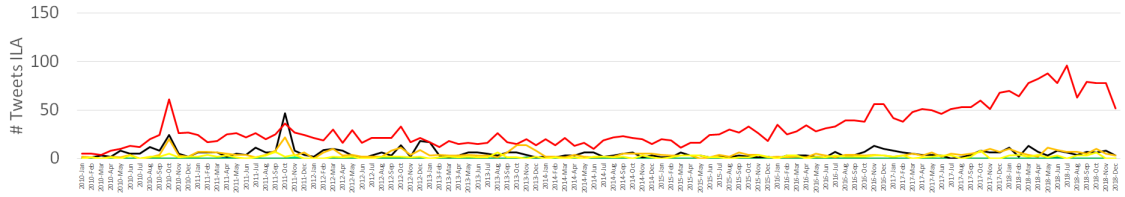

Figura 16: Comparação da distribuição de tweets que mencionam **psoríase**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

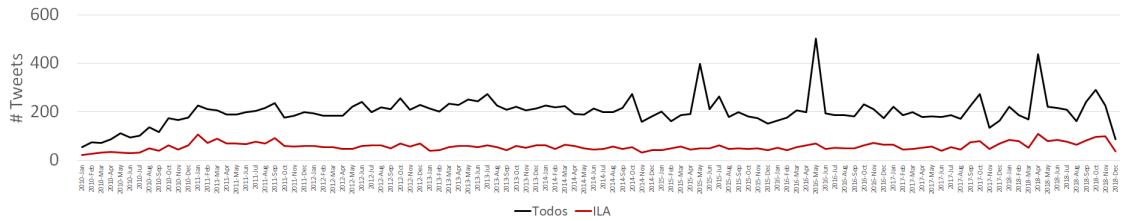

Figura 17: Comparação da distribuição de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**.

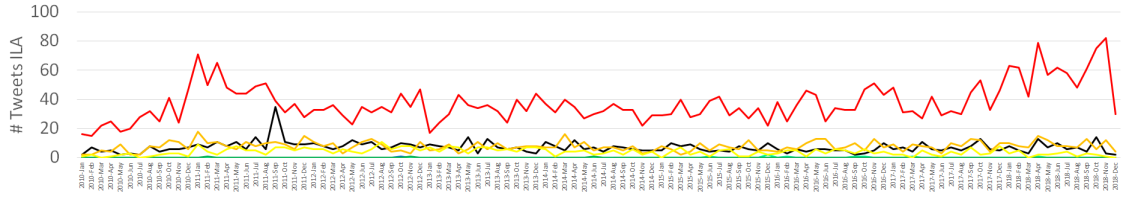

Figura 18: Comparação da distribuição de tweets que mencionam **urticária**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

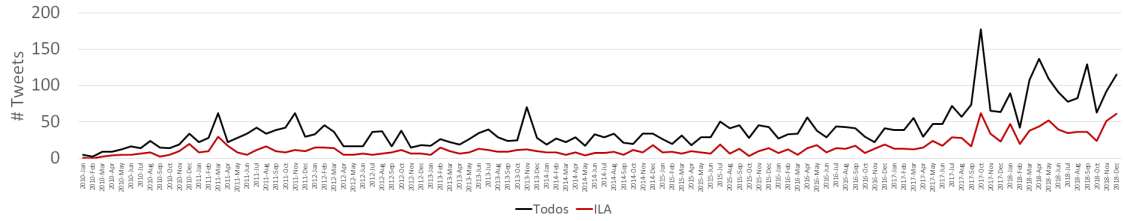

Figura 19: Comparação da distribuição de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**.

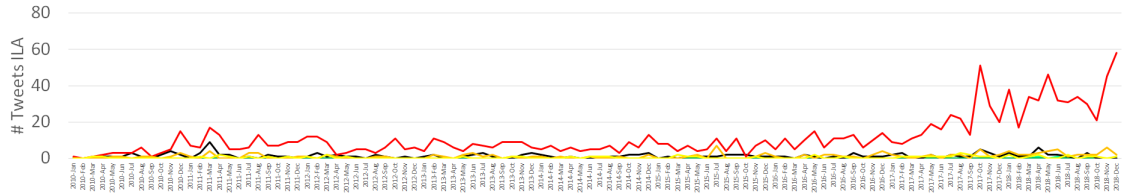

Figura 20: Comparação da distribuição de tweets que mencionam **dermatite atópica**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

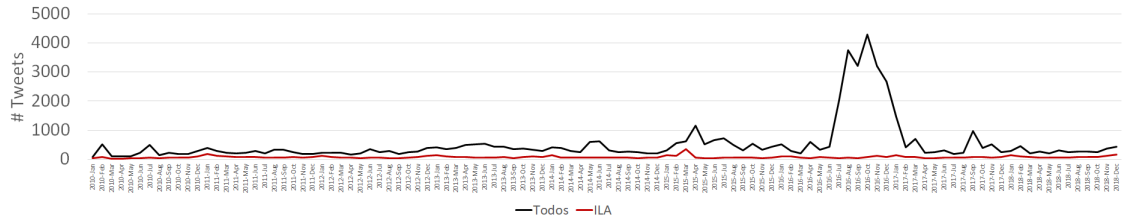

Figura 21: Comparação da distribuição de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**.

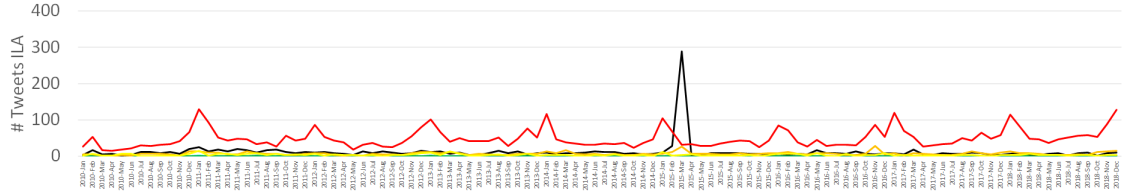

Figura 22: Comparação da distribuição de tweets que mencionam **vitiligo**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

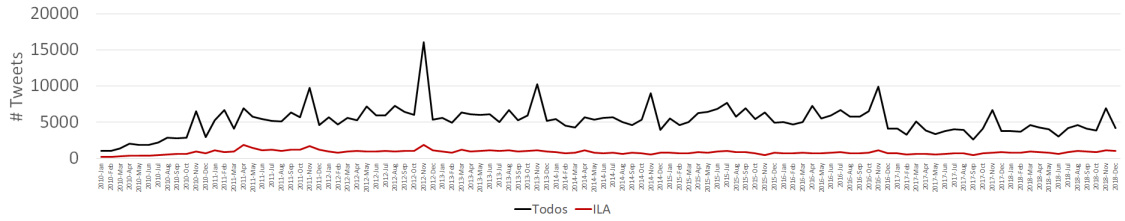

Figura 23: Comparação da distribuição de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**.

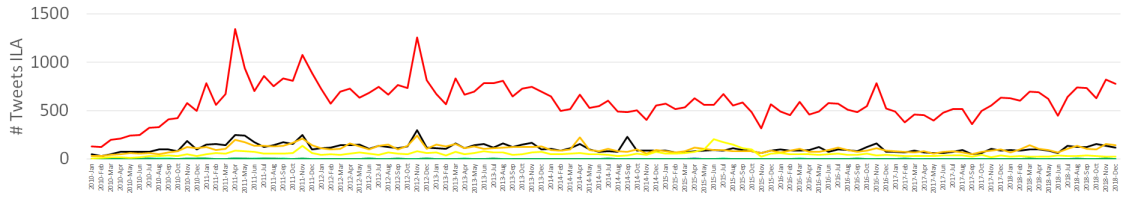

Figura 24: Comparação da distribuição de tweets que mencionam **diabetes**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

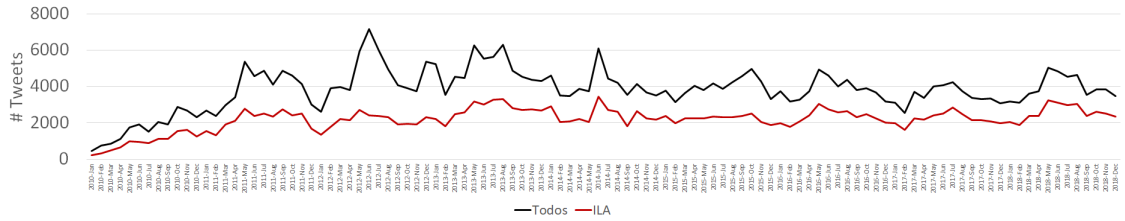

Figura 25: Comparação da distribuição de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**.

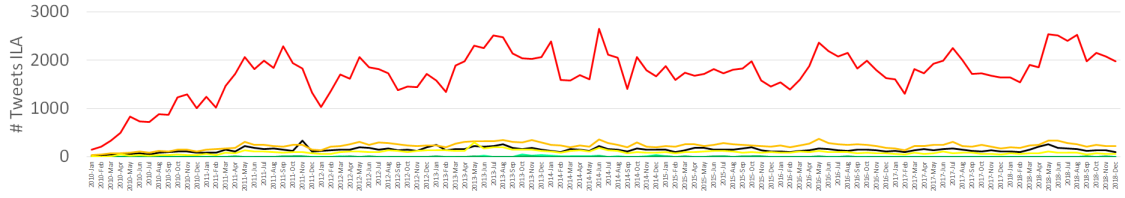

Figura 26: Comparação da distribuição de tweets que mencionam **asma**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**.

Lista de Figuras

2.1 Modelo Circunflexo de Afeto, Russell (1980). 6

2.2 Modelo de representação de emoções adotado, baseado no Modelo circunflexo de Russel. 6

2.3 SVM: Figura ilustrativa a duas dimensões. O "hiper-plano" é aqui ilustrado por uma recta e as margens por tracejado. 12

3.1 Espectro emocional do conjunto total de tweets respectivo a cada doença, por todos os quadrantes considerados e suas fronteiras. Resultados obtidos da análise de léxicos LEED e ANEW-PT. 28

3.2 Distribuição temporal de tweets referentes a **Psoríase** entre 2010 e 2018. . 29

3.3 Distribuição temporal de tweets referentes a **Urticária** entre 2010 e 2018. . 29

3.4 Distribuição temporal de tweets referentes a **Dermatite Atópica** entre 2010 e 2018. 29

3.5 Distribuição temporal de tweets referentes a **Vitiligo** entre 2010 e 2018. . . 30

3.6 Distribuição temporal de tweets referentes a **Diabetes** entre 2010 e 2018. . 30

3.7 Distribuição temporal de tweets referentes a **Asma** entre 2010 e 2018. . . . 30

3.8 Comparação da distribuição temporal de tweets referentes às seis doenças abordadas entre 2010 e 2018. 31

3.9 Distribuição temporal de tweets referentes a **Dermatite Atópica** entre 2010 e 2018, localizados em **Portugal**. 32

4.1 Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino exclusivo com dados obtidos via **crowdsourcing** e filtragem realizada tendo em conta o **classificador binário (ILA vs não-ILA)**. 45

4.2 Comparação da distribuição do número mensal de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 46

4.3 Comparação da distribuição do número mensal de tweets que mencionam **psoríase**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 46

4.4 Comparação da distribuição do número mensal de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 47

4.5 Comparação da distribuição do número mensal de tweets que mencionam **urticária**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 47

4.6 Comparação da distribuição do número mensal de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 47

4.7 Comparação da distribuição do número mensal de tweets que mencionam **dermatite atópica**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 48

4.8 Comparação da distribuição do número mensal de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 48

4.9 Comparação da distribuição do número mensal de tweets que mencionam **vitiligo**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 48

4.10 Comparação da distribuição do número mensal de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 49

4.11 Comparação da distribuição do número mensal de tweets que mencionam **diabetes**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 49

4.12 Comparação da distribuição do número mensal de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir do **classificador binário (ILA vs não-ILA)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. . . . 49

4.13 Comparação da distribuição do número mensal de tweets que mencionam **asma**, identificados como **ILA** a partir do **classificador binário (ILA vs não-ILA)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 50

4.14 Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino exclusivo com dados obtidos via **crowdsourcing** e filtragem realizada tendo em conta os **três classificadores (Autor, Sentido, Doente)**. 54

4.15 Comparação da distribuição de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. . . . 55

4.16 Comparação da distribuição de tweets que mencionam **psoríase**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 55

4.17 Comparação da distribuição de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. . . . 56

4.18 Comparação da distribuição de tweets que mencionam **urticária**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 56

4.19 Comparação da distribuição de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 56

4.20 Comparação da distribuição de tweets que mencionam **dermatite atópica**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 57

4.21 Comparação da distribuição de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. . . . 57

4.22 Comparação da distribuição de tweets que mencionam **vitiligo**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 57

4.23 Comparação da distribuição de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. . . . 58

4.24 Comparação da distribuição de tweets que mencionam **diabetes**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 58

4.25 Comparação da distribuição de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. . . . 58

4.26 Comparação da distribuição de tweets que mencionam **asma**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino exclusivo aos dados anotados em **crowdsourcing**. 59

5.1 Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino recorrendo **Active Learning** e filtragem realizada tendo em conta o **classificador binário (ILA vs não-ILA)**. 69

5.2 Comparação da distribuição de tweets que mencionam **psoríase**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **Active learning**. 70

5.3 Comparação da distribuição de tweets que mencionam **vitiligo**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referencial emocional adotado. Conjunto de treino complementado com **Active learning**. 70

5.4 Comparação da distribuição de tweets que mencionam **diabetes**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **Active learning**. 71

5.5 Comparação da distribuição de tweets que mencionam **asma**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **Active learning**. 71

5.6 Espectro emocional dos subconjuntos de tweets de cada doença por todos os quadrantes considerados e suas fronteiras. Resultados obtidos do treino recorrendo **Active Learning** e filtragem realizada tendo em conta os **três classificadores (Autor, Sentido, Doente)**. 74

1 Distribuição temporal de tweets referentes a **Psoríase** entre 2010 e 2018, localizados em **Portugal**. 86

2 Distribuição temporal de tweets referentes a **Urticária** entre 2010 e 2018, localizados em **Portugal**. 86

3 Distribuição temporal de tweets referentes a **Vitiligo** entre 2010 e 2018, localizados em **Portugal**. 86

4 Distribuição temporal de tweets referentes a **Diabetes** entre 2010 e 2018, localizados em **Portugal**. 87

5 Distribuição temporal de tweets referentes a **Asma** entre 2010 e 2018, localizados em **Portugal**. 87

6 Comparação da distribuição temporal de tweets referentes às seis doenças abordadas entre 2010 e 2018, localizados em **Portugal**. 87

7 Comparação da distribuição de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**. 88

8 Comparação da distribuição de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**. 88

9 Comparação da distribuição de tweets que mencionam **urticaria**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 88

10 Comparação da distribuição de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**. 88

11 Comparação da distribuição de tweets que mencionam **dermatite atópica**, identificados como **ILA** pelo **classificador binário** pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 89

12 Comparação da distribuição de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**. 89

13 Comparação da distribuição de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**. 89

14 Comparação da distribuição de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) pelo **classificador binário**. Conjunto de treino complementado com **active learning**. 89

15 Comparação da distribuição de tweets que mencionam **psoríase**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**. 90

16 Comparação da distribuição de tweets que mencionam **psoríase**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 90

17 Comparação da distribuição de tweets que mencionam **urticária**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**. 90

18 Comparação da distribuição de tweets que mencionam **urticária**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 90

19 Comparação da distribuição de tweets que mencionam **dermatite atópica**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**. 91

20 Comparação da distribuição de tweets que mencionam **dermatite atópica**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. . . 91

21 Comparação da distribuição de tweets que mencionam **vitiligo**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**. 91

22 Comparação da distribuição de tweets que mencionam **vitiligo**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 91

23 Comparação da distribuição de tweets que mencionam **diabetes**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**. 92

24 Comparação da distribuição de tweets que mencionam **diabetes**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 92

25 Comparação da distribuição de tweets que mencionam **asma**, entre o conjunto total e o conjunto filtrado: tweets identificados como **ILA** (a vermelho) a partir dos **três classificadores (Autor, Sentido, Doente)**. Conjunto de treino complementado com **active learning**. 92

26 Comparação da distribuição de tweets que mencionam **asma**, identificados como **ILA** a partir dos **três classificadores (Autor, Sentido, Doente)**, pelos vários quadrantes e fronteiras do referêncial emocional adoptado. Conjunto de treino complementado com **active learning**. 92

Lista de Tabelas

2.1 Exemplo ilustrativo da aplicação do léxico para classificar um texto que inclui os emojis listados, calculando a média. Os valores estão normalizados entre 0 e 1. O valor 0.5 separa a valência negativa (<0.5) da positiva (>0.5), bem como a ativação baixa (<0.5) da alta (>0.5). 7

2.2 Seis análises do estudo linguístico, segundo Jurafsky and Martin (2009). . . 8

2.3 Notação usada na tabela 2.4. 13

2.4 Adaptação das métricas *Precisão*, *Recall* e *F1* ao problema multi-*label*. Adaptado do site da biblioteca `Scikit-Learn`. 13

3.1 Variantes das referências às doenças, reduzidas a letra minúscula. 21

3.2 Número total de tweets obtidos por doença e número de tweets que não a mencionam. 21

3.3 20 tópicos obtidos com LDA a partir de **todos os tweets** das quatro psicodermatoses. 24

3.4 Número de tweets provenientes de Portugal, Brasil e local indefinido. 31

3.5 Percentagem de tweets provenientes de Portugal, Brasil e local indefinido. . 32

4.1 Distribuição dos dados anotados via **crowdsourcing** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores. 36

4.2 Distribuição em **percentagem** dos dados anotados via **crowdsourcing** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores. 36

4.3 Distribuição dos dados anotados via **crowdsourcing** pelos vários níveis de valência e ativação (1 a 5), e nas classes geradas posteriormente (simplificação para três labels possíveis: Negativo (-), Origem (0), Positivo(+)). . . 36

4.4 Distribuição em **percentagem** dos dados anotados via **crowdsourcing** pelos vários níveis de **valência** e **ativação** (1 a 5), e nas classes geradas posteriormente (simplificação para três labels possíveis: Negativo (-), Origem (0), Positivo (+)). 37

4.5 Distribuição dos dados anotados via **crowdsourcing** pelos vários **quadrantes** e fronteiras do referêncial emocional adoptado. 38

4.6 Distribuição em **percentagem** dos dados anotados via **crowdsourcing** pelos vários **quadrantes** e fronteiras do referencial emocional adotado. . . 38

4.7 Melhor combinação de características para cada problema de classificação. Conjunto de treino: *crowdsourcing. Class. Binário* corresponde ao classificador binário que separa os tweets em cujo autor é o doente e fala no sentido literal, dos restantes. 39

4.8 Performance dos classificadores selecionados a partir de *10-fold cross-validation.* Precision-macro, Recall-macro, F1-macro e F1-weighted. Conjunto de treino: **crowdsourcing**. 40

4.9 Distribuição de tweets que mencionam cada doença entre várias classes. . . 41

4.10 Distribuição percentual de tweets que mencionam cada doença entre várias classes. 41

4.11 Comparação de resultados de filtragem de tweets ILA, por duas modalidades: Classificadores ASD (cruzamento de resultados dos classificadores individuais de Autor, Sentido e Doente) e Classificador Binário. 41

4.12 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelo **classificador binário** treinado com os dados de **crowdsourcing**. 42

4.13 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelos três classificadores **(Autor, Sentido, Doente)** treinados com os dados de **crowdsourcing**. 51

5.1 Active Learning - exemplo ilustrativo: cálculo da incerteza (menor confiança na classificação), critério para recomendar instância para anotação. Neste caso, o Tweet B seria o escolhido. 62

5.2 Distribuição dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores. 63

5.3 Distribuição em **percentagem** dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelas várias classes (Autor, Sentido, Doente) e pela binarização dos dados anteriores. 63

5.4 Distribuição dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelos vários níveis de valência e ativação (1 a 5), e nas classes geradas posteriormente (simplificação para três labels possíveis: Negativo (-), Origem (0), Positivo (+). 63

5.5 Distribuição em **percentagem** da reunião dos dados anotados via crowdsourcing com as anotações com recurso a **active learning** pelos vários níveis de **valência** e **ativação** (1 a 5). 63

5.6 Distribuição em **percentagem** da reunião dos dados anotados via crowdsourcing com as anotações com recurso a **active learning** pelos vários níveis de **valência** e **ativação** simplificados: Negativo (-), Origem (0), Positivo (+). 64

5.7 Distribuição dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelos vários **quadrantes** e fronteiras do referêncial emocional adoptado. A identificação do tweet num quadrante foi feita com base nas anotações de valência e ativação (1 a 5). 64

5.8 Distribuição em **percentagem** dos dados anotados via crowdsourcing e anotações pessoas com recurso a **active learning** pelos vários **quadrantes** e fronteiras do referêncial emocional adoptado. A identificação do tweet num quadrante foi feita com base nas anotações de valência e ativação (1 a 5). 64

5.9 Classificadores adoptados. Conjunto de treino: *crowdsourcing* e *Active Learning*. *Class. Binário* corresponde ao classificador binário que separa tweets cujo autor é o doente e fala no sentido literal, dos restantes. 64

5.10 Performance dos classificadores a partir de *10-fold cross-validation*. Precision-macro, Recall-macro, F1-macro e F1-weighted. Conjunto de treino: crowdsourcing e **active learning**. 65

5.11 Distribuição de tweets que mencionam cada doença entre várias classes. . . 65

5.12 Distribuição percentual de tweets que mencionam cada doença entre várias classes. 65

5.13 Distribuição percentual de tweets que mencionam cada doença entre várias classes. 65

5.14 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelo **classificador binário** treinado com os dados de **crowdsourcing e active learning**. 66

5.15 20 tópicos obtidos com LDA a partir dos **tweets ILA** das quatro psicodermatoses, filtrados pelos três classificadores **(Autor, Sentido, Doente)** treinados com os dados de **crowdsourcing e active learning**. 72

6.1 Popularidade de cada doença, relativa aos subconjuntos ILA, filtrados por classificador Binário (Bin.) e três classificadores (ASD), treinados sobre o conjunto exclusivo a crowdsourcing (CS) ou complementado com active learning (AL). 78

6.2 Comparação dos resultados das quatro modalidades de classificação estudadas nos capítulos 4 (CS: crowdsourcing) e 5 (AL: active learning), com os resultados originais do capítulo 3. '#' designa 'número'. Entre parênteses está incluído o número de tópicos confusos, eventualmente compatíveis com tweets ILA. Dois agrupamentos de colunas: resultados observados pela análise de tópicos e eventos detetados na distribuição temporal relativa a vitiligo. . 79