Mestrado em Engenharia Informática

Estágio

Relatório Final

SoFly: meet the social engineering

José Miguel Malaca Cardoso

jmmcar@student.dei.uc.pt

Orientador do DEI:

Carlos Manuel Mira da Fonseca

cmfonsec@dei.uc.pt

Orientador da Empresa:

Virgílio Manuel Raposo Esteves

virgilio@broadscope.eu

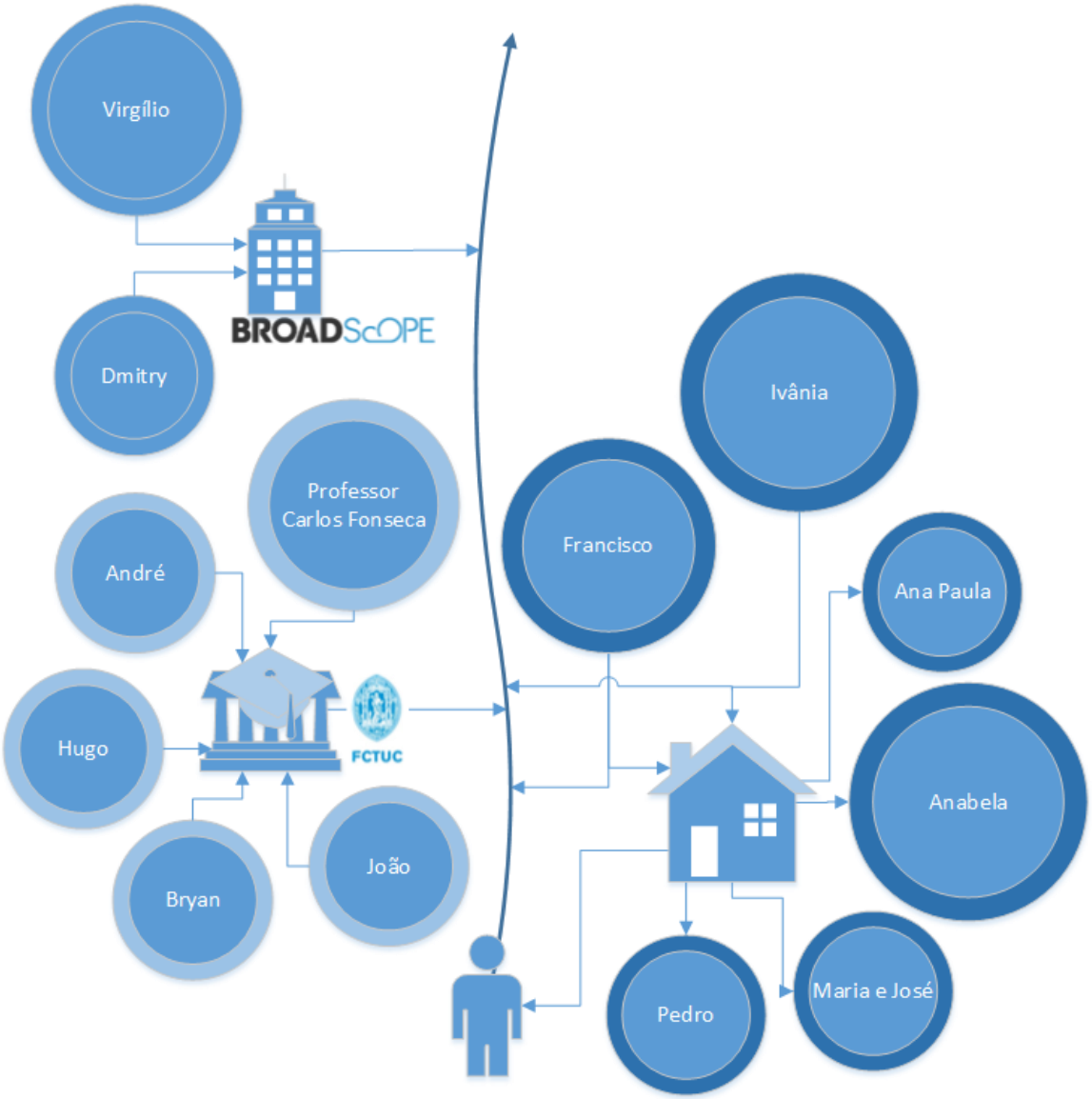
Data: 2 de Setembro de 2014

FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Aos Malacas e Marques da minha vida, com todo o carinho.

Agradecimentos

Este espaço é dedicado àqueles que contribuíram de alguma forma para o meu percurso académico que culmina na realização deste trabalho. Família, amigos, colegas, professor, a todos, expresso aqui o meu agradecimento sincero…



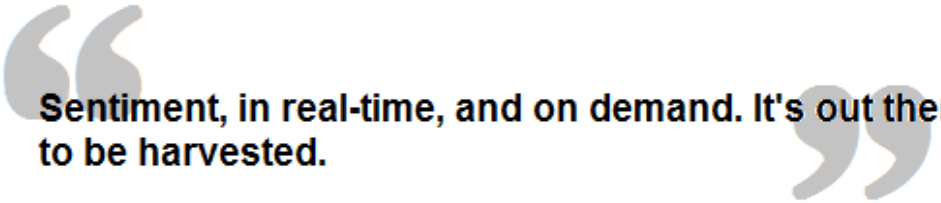
Obrigado!

Resumo

Com a crescente afluência do público em geral às redes sociais há um vasto leque de informação a ser explorado. A facilidade com que a tecnologia permite às pessoas comunicar, e a enorme vontade que elas têm de partilhar muitos aspectos das suas vidas, originaram um enorme mundo de informação. Sendo toda esta informação pública e disponível a quem a quiser analisar, como se poderá descobrir a opinião ou o sentimento que uma dada população expressa ou expressou em relação a um determinado assunto? Estes problemas conhecidos por "*Opinion Mining*" e "*Sentiment Analysis*", têm vindo a ser abordados pela comunidade ciêncitifica através de métodos de *Machine Learning* (ML) e de *Natural Language Processing* (NLP). Neste estágio, realizado na empresa BroadScope, foram analisadas, implementadas e avaliadas diversas técnicas de análise de sentimento em textos no formato de *microblog* de onde se obteram informações relevantes para a classificação de sentimento, como também, posteriormente, bons resultados de classificação de textos nunca vistos pelo sistema.

Palavras-Chave

"Text Mining", "Social Mining", "Sentiment Analysis", "Machine Learning", "Natural Language Processing", "Twitter"



Kim Davis @ Feelings: The NYC Sentiment Symposium [1]

[1] Http://www.saasintheenterprise.com/author.asp?section_id=3292&doc_id=272067

Índice

Índice de Figuras

Índice de Tabelas

Índice de Gráficos

Índice de Pseudocódigos

Capítulo 1

Introdução

Segundo *Pang* e *Lee* [1], o ano de 2001 marcou o início de uma consciência generalizada para os problemas de pesquisa e oportunidades que trariam as áreas de análise de sentimento e *Opinion Mining*. *Opinion Mining* define-se como a identificação da opinião do público acerca de um dado produto a partir de críticas, mensagens de *blogs* e comentários através do processamento desses textos. A análise dos textos pode ser realizada por métodos simples, tal como a verificação da existência de palavras pré-classificadas como positivas ou negativas, mas também pode envolver uma análise mais profunda da estrutura gramatical dos textos na tentativa de levar a uma melhor compreensão de aspectos relevantes desses textos. O processamento de linguagem natural (*Natural Language Processing, NLP*) é uma área que liga a inteligência artificial e a linguística com o objectivo de levar os computadores a realizar tarefas que envolvem a compreensão e/ou geração deste tipo de linguagem. Para pessoas não ligadas ao ramo linguístico, este processo pode ser considerado pesado e penoso, havendo grande interesse em outros métodos capazes de extrair a informação desejada sem que seja necessário um estudo aprofundado da língua.

De facto, existem algoritmos da área de Machine Learning (ML) que permitem identificar os possíveis sentimentos expressos em textos usando sobretudo dados estatísticos, e não apenas linguísticos, para extrair a informação desejada, transformando-se assim o problema linguístico num problema de classificação.

Com a popularização de análise de conteúdos e de extração de informação de textos, nasce o termo *Sentiment Analysis*. A Análise de Sentimento é um processo que tenta atribuir correctamente uma polaridade aos textos que recebe à entrada, tendo sido utilizada com sucesso em vários domínios. O principal tipo de texto, e o estudado neste trabalho, tem as características de um serviço designado por *Microblogging*. *Microblogging* é um serviço *web* que possibilita ao seu assinante publicar mensagens curtas para outros assinantes. O mais popular tem por nome *Twitter* e, ao longo do tempo, foi criando uma rede de milhões de assinantes, tendo-se tornado uma das principais redes sociais dos dias de hoje. Muitas pessoas influentes a utilizam, bem como marcas e empresas, com a principal intenção de chegar ao maior número possível de pessoas e as influenciar de alguma maneira. Também estas pessoas, mais anónimas, têm as suas opiniões e as exprimem publicando nas mesmas redes. É no contexto de todas estas opiniões e reações que a Análise de Sentimento encontra maior relevância. Ao identificar a polaridade dos textos está-se a obter informação sobre o

feedback, a opinião ou o sentimento que as pessoas exprimem sobre um dado assunto através das redes sociais. Esta informação permite às empresas conhecer a reação de toda uma população aos seus produtos, tanto em tempo real como relativamente a um determinado intervalo de tempo, tornando-se assim informação de grande valor.

1.1. Motivação

Este trabalho foi desenvolvido no âmbito do estágio curricular do Mestrado em engenharia Informática realizado na *BroadScope*, empresa que iniciou atividade em 2012 e se encontra sediada na Incubadora do Instituto Pedro Nunes, em Coimbra. A *BroadScope* é uma empresa multidisciplinar com especial enfoque em tecnologias *Microsoft*. Constituída por uma jovem equipa, desenvolve soluções *web*, distribuídas, móveis e *desktop*, tendo já uma comprovada qualidade reconhecida pela atribuição de prémios e certificações internacionais, tais como *Microsoft Certified Professional*, *Microsoft Certified Technology Specialist* e *Microsoft Most Valuable Professional*, como também pelo leque dos seus clientes tais como a *Microsoft*, *Indra*, *Leadership Business Consulting*, *Qolpac*.

Apesar da *BroadScope* ser uma consultora tecnológica, já deu origem a uma *spin-off* denominada *SoFly*, nome este que é partilhado pelo produto no qual assenta. O *SoFly* é uma plataforma de *second-screen* que altera a forma como o indivíduo consome, experimenta e interage com conteúdos de media. O produto destina-se a dois segmentos distintos, público em geral e empresas, tendo a vertente empresarial a designação de *SoFly Analytics*.

O *SoFly Analytics* recorre às redes sociais para recolher um conjunto lato de informação para ser posteriormente processada. Uma parte importante deste processamento passa pela identificação de empresas, produtos, serviços e pessoas, bem como pela identificação do sentimento associado à informação recolhida. É neste contexto que este estágio se insere.

1.2. Objetivos

O principal objetivo do estágio consistiu no estudo e implementação de métodos e processos de análise de sentimento culminando no desenvolvimento de um módulo capaz de atribuir a classificação de positivo, neutro ou negativo a textos no formato de Microblog (mais concretamente, *tweets*). Para tal, foram implementas várias possibilidades para permitir a comparação de resultados e opções através de estudos e análises, que se encontram documentadas no presente relatório. O estágio visou ainda o ensaio do módulo desenvolvido no estudo de um caso real.

1.3. Contribuições

O estágio que o aluno realizou durante um ano letivo, contribuiu principalmente para a introdução deste num ambiente empresarial com o desenvolvimento de um sistema desejado pela empresa hospedeira, sendo que no âmbito da empresa, este trabalho realizado pelo aluno contribui para a incorporação de novos conhecimentos.

Este projeto apresenta-se como extremamente importante para a empresa e possui um cariz crucial. A possibilidade de dotar o sistema SoFly Analytics de análise em tempo real de informação obtida de redes sociais e de identificação de sentimentos expressos nessa mesma informação, de uma forma escalável, repetível e de baixo custo representa um fator diferenciador entre o produto da empresa e os produtos disponíveis no mercado. Durante o decorrer do estágio, foi possível comprovar já a eficiência de alguns dos processos devido à sua introdução faseada no produto aquando da realização de um caso de estudo com um cliente real, tendo sido obtidos resultados e *feedback* do cliente extremamente positivos.

1.4. Estrutura do Documento

O presente documento encontra-se dividido em várias secções a fim de facilitar a sua leitura e compreensão. São abordados primeiramente as bases do tema análise de sentimento. Em particular são revistos de modo conciso e direto os problemas e soluções existentes, colocando o leitor a par dos vários assuntos que constituem esta análise. Posteriormente, é apresentado o trabalho desenvolvido, bem como os estudos, análises e implementações realizados durante todo o estágio para que, por fim, sejam apresentados os resultados e se retirem conclusões.

Capítulo 2

Análise de Sentimento

Um dos primeiros grandes trabalhos sobre a análise de sentimento foi realizado por *Pang* e *Lee* [1]. Este tornou-se numa grande base para qualquer estudo nos temas, *Opinion* ou *Sentiment Analysis*, levando a que na grande maioria dos trabalhos existentes na literatura se encontre, de uma forma ou de outra, uma referência a este trabalho. É um trabalho completo sobre a reunião de informação orientada a textos de opinião. Mais concretamente, discute quais serão nas áreas de *Opinion* ou *Review-search* as possibilidades de aplicação, desafios, abordagens e implicações da utilização de classificadores para a extração de informação de textos. *Pang* e *Lee* [2] mostraram que a classificação multi-classe conduz a bons resultados. Esta classificação passa por duas abordagens, a supervisionada e a não supervisionada. A abordagem supervisionada é baseada em algoritmos de aprendizagem (*machine learning*). Recorrendo a uma coleção de dados iniciais pré-classificados treina-se um classificador para que este preveja a classificação de novos dados. A abordagem não supervisionada, também chamada de abordagem de orientação semântica, tenta medir se um texto está inclinado a um sentimento positivo ou negativo através de uma análise lexical e sintáctica recorrendo para isto a métodos da área do processamento da linguagem natural (*Natural Language Processing, NLP*). *Pang* e *Lee* [2] e *Chaovalit* e *Zhou* [3] compararam estas duas abordagens aos textos de opinião, tentando classificar revisões de filmes como positivas ou negativas, definindo assim o problema da Polaridade. Pode ser também considerado um outro grupo de textos, o neutro, mas com este surge um outro problema, o da Subjetividade. Este representa a separação entre textos com algum tipo de sentimento (chamados de Subjetivos) e aqueles sem qualquer tipo de sentimento (Objetivos ou Neutros). Olhando de forma geral para vários trabalhos, todos se focam principalmente na abordagem supervisionada, e para a executarem seguiram um conjunto de passos de forma mais ou menos aprofundada:

a) Aquisição de dados,
b) Pré-processamento dos dados,
c) Seleção de *features*,
d) Escolha do tipo de classificador,
e) Treino e validação do classificador.

Estes passos definem o processo para a criação de um classificador baseado em algoritmos de aprendizagem. Deverá existir um conjunto de dados pré-classificados para treino e validação do classificador. Deverá ser realizado algum tipo de pré-

processamento para preparar os dados e para permitir selecionar as *features* em que o classificador baseará a sua aprendizagem.

Este capítulo apresenta um estudo da literatura sobre o tema Análise de Sentimento em *Tweets*. Foram lidos e estudados vários aspectos nesta área, com o intuito de aprofundar conhecimento e extrair informação relevante para o trabalho. São introduzidos vários pontos sobre a análise de sentimento começando pela aquisição de dados na secção 2.1.. Nas secções 2.2. e 2.3. são consideradas as abordagens supervisionada e não supervisionada, respectivamente, seguindo-se a avaliação de desempenho na secção 2.5. e terminado com algumas ferramentas disponiveis sobre a análise de sentimento em *tweets*, secção 2.6.. Finalmente, é feita uma apreciação crítica na secção 2.7.

2.1. Aquisição de Dados Pré-Classificados

Os dados a que se recorrerá para a extração de informação serão textos com uma dimensão reduzida e com características muito próprias. Provêm de um serviço existente na internet com o nome de *Twitter*[2]*,* que representa uma das principais redes socias do mundo actual. Este baseia-se no conceito de *Microblogging*. *Microblog* é apresentado por *Adam* e *Alan* [4] como um formato de texto de tamanho reduzido. No caso dos *tweets,* existe o limite de cento e quarenta caracteres.

A aquisição de dados é realizada pela maioria dos trabalhos revistos através da *Application Programming Interface* (API) do *Twitter.* Recorrendo a pedidos (*querys*) recolhem-se *tweets* da sua *feed* sem qualquer tipo de restrição, mas nem todos os autores recolhem os dados desta forma. Pode ser especificado, por exemplo, o assunto (*target*) a que os *tweets* devem estar ligados, como um produto, um serviço ou uma pessoa, ou que tipos de *emoticons* os *tweets* devem conter, positivos ":-)" ou negativos ":-(". Também é possível recolher os *tweets* através de outros serviços disponiveis na internet, como no trabalho de *Barbosa* e *Feng* [5] onde é descrito que recorreram ao *Twendz*[3]*, a*o *Twitter Sentiment*[4] e ao *TweetFeel*[5] para reunir o seu corpus de *tweets*. Com os dados obtidos, a sua pré-classificação é realizada seguindo algumas ideias chave.

Primeiramente definir-se-á sentimento no *twitter.* Define-se que a "análise de sentimento é a tarefa de identificar opiniões, emoções e avaliações positivas ou negativas" [4], e que a polaridade estará ligada a "um sentimento positivo ou negativo" [6] deixando de

[2] Https://twitter.com/
[3] Http://twendz.waggeneredstrom.com/
[4] Http://www.sentiment140.com/
[5] Http://www.tweetfeel.com/

lado o aspecto da neutralidade. Nos trabalhos que incluíram este aspecto alguns fizeram a atribuição manualmente, enquanto outros definem que *tweets* neutros são textos que poderiam fazer parte de um título de um jornal [7]. Esta definição, apesar de muito discutível, será uma boa forma de definir a neutralidade em *tweets*, levando a recolha deste tipo de *tweets* para contas de jornais de renome, como o *New York Times* ou o *Washington Post.* No caso da utilização dos serviços disponíveis na internet, estes já apresentam os seus *tweets* classificados. Recorrendo à API do *Twitter* e a *emoticons* positivos e negativos*,* recolhem-se os *tweets* e atribui-se a respectiva polaridade. *Tweets* com um *emoticon* pré-classificado como positivo são automaticamente classificados como *tweets* positivos. Acontece o mesmo para o caso contrário, o negativo. Sabendo-se que o uso da API do *Twitter* poderá devolver muitos *tweets* que não contêm qualquer informação sentimental, ou que, mesmo manualmente, não se lhes conseguirá atribuir uma polaridade, foi definido no trabalho de *Agarwal et al*. [8] que estes deveriam ser classificados como "*junk"* e excluídos. É importante referir que esta pré-classificação sobre a polaridade de um *tweet* o representará como um todo. O *emoticon* encontrado representará todo o *tweet* e consequentemente todas as suas palavras. É importante esta definição, visto que existe sempre a possibilidade de um *tweet* expressar informação sobre vários assuntos, mas este caso não será abordado neste trabalho.

2.2. Análise Não supervisionada

Como *Pang* e *Lee* [2] descreveram, uma abordagem ao problema da classificação é a de orientação semântica. Também chamada de Análise não Supervisionada, tem como objetivo a implementação de um sistema capaz de classificar o sentimento expresso nos textos recorrendo a métodos de NLP. Um modelo para este tipo de análise não supervisionada passa pelo descrito na Figura 3.

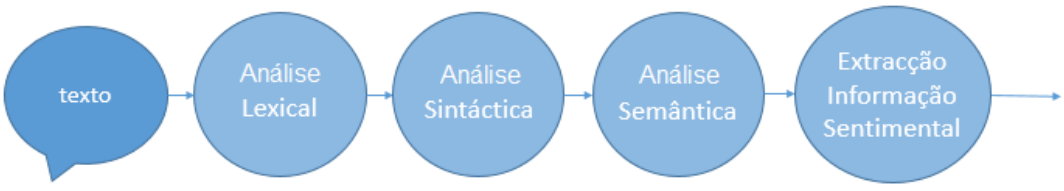


Figura 1: Modelo não supervisionado para a extração de dados de textos

Neste modelo realiza-se uma completa análise linguística aos textos. Como descrito, a recolha de informação é realizada partindo de três tipos de análise que podem ser estudadas e implementadas. A Lexical, que se focará em retirar informação sobre as

próprias palavras ou *tokens* existentes no texto. A Sintática, que representa uma análise gramatical. Verifica-se entre as várias palavras e *tokens* do texto as suas classes gramaticais e as ligações entre umas e outras. Por fim a Semântica que recolhe informação sobre os temas, tópicos e assuntos abordados no texto.

Sobre a extração de sentimento em *tweets* foram encontrados os trabalhos de *Nasukawa* e *Yi* [9] e *Jiang et al.* [10], onde se utilizam métodos da análise lexical e sintática na tentativa de extrair sentimentos positivos ou negativos de textos. Estes sentimentos foram relacionados com assuntos específicos e o porquê prende-se com o objetivo destes. Querendo avaliar textos e simplificando o problema, utilizam-se textos bem escritos. Descrevem principalmente revisões sobre filmes como *Pang* e *Lee* [2] ou *Chaovalit and Zhou* [3], onde se desejou encontrar os assuntos dos textos e assim tentar definir qual o sentimento que está expresso sobre este. Ou seja, procuraram-se relações gramaticais entre palavras para se definir o sentimento. Focaram-se sobre uma análise sintática aos textos onde recorrem a classes gramaticais que representam o sentimento atribuído a nomes e acções. Por exemplo, recorrendo:

a) Ao adjectivo "*good*". Se "*good*" tiver a notação de positivo, uma pequena frase como "*good product*" obterá um sentimento positivo. Um sentimento positivo atribuído por este adjetivo à palavra que lhe está ligada, o nome "*product*",
b) Ao advérbio "*beautifully*". Estes tendem a alterar o significado aos verbos a que estão ligados. Na frase "*play beautifully*", o verbo "*play*" herdará o sentimento que o advérbio "*beautifully*" contém.
c) No caso dos verbos, o sentimento irá depender na relação com os seus argumentos. No exemplo, "XXX *beats* YYY", verifica-se que o verbo está atribuir um sentimento positivo ao sujeito e um negativo ao objecto.

Para este tipo de análise utiliza-se um método com o nome de *POS-Tagging*[6] para desambiguar expressões que poderão ser encontradas, através da atribuição de classes gramaticais às palavras. No caso da palavra "like", poderá existir um sentimento associado, mas apenas quando esta é um verbo, e não se for um adjectivo ou uma preposição. O sistema *Talent System* baseado na arquitetura *TEXTRACT* [11], o *Stanford Parser*[7] e o *OpenNLP Parser*[8] são ferramentas deste tipo Estas executam a análise sintáctica devolvendo as várias classes gramaticais atribuídas às palavras como também as ligações existentes entre elas.

Uma explicação para que esta abordagem não supervisionada não obtenha resultados muito bons, como *Chaovalit* e *Zhou* [3] ou *Pang* e *Lee* [2] descrevem, é que este

[6] Http://tinyurl.com/pjc75sc
[7] Http://nlp.stanford.edu/software/lex-parser.shtml
[8] Http://opennlp.apache.org/cgi-bin/download.cgi

processo não tem em conta o tópico a analisar. Todas as frases são processadas, mesmo as que não contêm qualquer informação relevante sobre o assunto pelo qual foram recolhidas. Assim poderá ser alterado o sentimento que seria atribuído aos textos. A classificação de *tweets* poderá passar também por este processo. Em vez de considerar um *tweet* como um todo, começar por extrair informação sobre o tópico ou até os seus subtópicos, se existirem.

2.3. Análise Supervisionada

Na literatura sobre a Análise de Sentimento em *tweets*, a abordagem supervisionada é a mais utilizada. Tal como se viu na secção anterior sobre a análise não supervisionada, a análise linguística de textos é muito complexa e trabalhosa, para além de estar fortemente dependente da língua dos textos. Assim será natural transformar o problema num de classificação. Recorrer a classificadores, a algoritmos de aprendizagem, é a forma menos complexa de abordar e solucionar o problema.

2.3.1. Pré-Processamento

Os dados para este trabalho são *tweets* com características próprias. Essas características devem ser analisadas para se perceber que informação é relevante para a solução do problema de classificação. Sendo os *tweets* textos de uma rede social, com um número de caracteres limitado, encontrar-se-ão formas simples e curtas das pessoas se expressarem. Além disso, o próprio *Twitter* fornece algumas formas específicas dos utilizadores comunicarem entre si, como por exemplo *hashtags, retweets, usernames*.

Assim, deverá existir um pré-processamento com o fim de preparar os dados para a sua classificação. Num primeiro passo executa-se uma análise lexical, em que se substituem certas palavras específicas dos *tweets* por pseudo-termos. Serão consideradas as seguintes categorias:

a) Hiperligações (*URLs*),
b) Palavras com o caracter arroba (@) na sua primeira posição (*Usernames*),
c) Palavras com o caracter cardinal (#) na sua primeira posição (*Hashtags*),
d) Palavras especiais do *Twitter* como “RT”, significando *retweet*,
e) Outras palavras escritas em maiúsculas (*Uppercase*).

Dos trabalhos lidos apenas *Go et al* [6] e *Agarwal et al.* [8] referem quais os pseudo-termos utilizados. *Usernames* foram substituídos por “USERNAME” e as hiperligações por “URL”. No trabalho de *Agarwal et al.* é referido que os *emoticons* são também substituídos, uns por “*POSITIVE*”, outros por “EXTREMELY POSITIVE”, e igualmente

para os *emoticons* negativos. Também é feita a substituição de negações como "*not*", "*no*" e "*never*" pelo pseudo-termo "NOT".

Além deste, também pode ser aplicado, o *POS-Tagging* (*Part-Of-Speech Tagging*). As classes gramaticais das palavras existentes num texto poderão ser bons indicadores da polaridade do texto e portanto não devem ser descartadas. Uma dada palavra será um nome, um verbo, um adjectivo, um pronome ou uma das trinta e sete classes apresentadas pelo trabalho *Penn Treebank* [12]. Estas classificações gramaticais podem ser posteriormente utilizadas através da substituição da palavra a que correspondem por "palavra_classe gramatical".

Em suma, as abordagens supervisionadas também utilizam técnicas de análise linguística na fase de pré-processamento, embora o processo de classificação posterior seja fundamentalmente diferente do das abordagens não supervisionadas.

2.3.2. Features

F*eatures* são particularidades, características ou elementos mensuráveis que se podem extrair dos textos, com vista a definição e separação das diferentes classes. As *features* selecionadas irão depender do problema a resolver e das classes a separar. No contexto do presente estudo colocam-se as seguintes questões:

i. Subjetividade. Que *features* permitirão separam melhor as classes de *tweets* Subjetivos e Objetivos?
ii. Polaridade, Que *features* permitirão separar melhor as classes de *tweets* Positivos e Negativos?

Poderão existir mais mas, para ambos os problemas, as *features* selecionadas estão maioritariamente relacionadas com três grandes grupos:

a) *Syntax-features*, que se referem às características das próprias palavras. Por exemplo, se elas se relacionam com a subjetividade [13], qual a sua polaridade[9] à priori, se existe alguma negação no texto e a que palavras essa negação diz respeito, invertendo assim a polaridade existente e a que classes gramaticais as palavras pertencem,
b) *Meta-features*, que representam dados que se encontram em *tweets,* tais como as *hashtags*, *retweets*, repostas (*replys*), hiperligações (*urls*), pontuação (principalmente exclamações ou interrogações), *emoticons* ou o uso de maiúsculas (*uppercase)*.
c) *N-grams*, que representam sequências de palavras que poderão ser utilizadas para a detecção de padrões na escrita. Considerando o N como

[9] Http://www.cs.pitt.edu/mpqa/

um valor de um a três, obtêm-se os chamados *unigrams*, *bigrams* e *trigrams* que serão utilizados na resolução do problema.

Na literatura a escolha das *features* parece simples. Encontram-se estudos que utilizam apenas *N-grams* ([14]), outros que utilizam *N-grams* e *meta-features* ([5]), ou ainda *meta e syntax-features* ([7]). Existindo a possibilidade de ocorrerem más classificações devido ao excesso, ou à falta, de *features* representativas das classes, existem métodos analíticos com o objetivo de identificar conjuntos de *features* que realmente representam informação discriminativa das classes. Um destes métodos passa pelo simples cálculo da frequência de cada *feature*. É um método [15] estatístico simples que considera o número de vezes que uma dada palavra ocorre nos textos de uma dada classe. Ou, alternativamente, o número de textos de uma dada classe em que essa palavra ocorre.

Existem outros métodos de seleção de *features* que poderão ser utilizados. O cálculo da informação mútua [15] mede a quantidade de informação com que a presença/ausência de uma dada *feature* contribui para uma decisão de classificação correcta. Para se calcular a informação mútua entre as várias *features* e classes, apresenta-se a expressão seguinte:

$$IM(U,C) = \sum_{e_f \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_f, C = e_c)\, log_2 \frac{P(U = e_f, C = e_c)}{P(U = e_f) P(C = e_c)}$$

Com U representando existir ou não (1 ou 0) uma dada *feature* no *tweet*, e C o *tweet* pertencer a uma dada classe, calcular-se-á a informação mútua referente sobre cada *feature,* relacionando estas com cada uma das classes de *tweets*. O segundo método tem por nome *Chi Square* [15] e é um outro método popular para a seleção de *features.* Usualmente, em estatística, é aplicado para testar a independência de dois eventos independentes. Formalmente apresenta-se através da expressão seguinte:

$$X^2(D,f,c) = \sum_{e_f \in \{1,0\}} \sum_{e_c \in \{1,0\}} \frac{\left(N_{e_f e_c} - E_{e_f e_c}\right)^2}{E_{e_f e_c}}$$

Onde D representa os dados, *f* a *feature* e c uma classe. *N* corresponderá à frequência observada e *E* à esperada.

Para o cálculo da utilidade de uma dada *feature* sobre uma classe de *tweets* quer recorrendo ao cálculo sobre a informação mútua, quer ao *Chi-Square*, apresenta-se o seguinte Pseudocódigo 1.

```
Selecão_Features (Tweets, Classe):
        Vocabulário = Extrair_Vocabulário (Tweets)
        Resultados = []
        Para cada Feature no Vocabulário:
                Valor (feature, classe) = Função_Calculo (Tweets, Feature, Classe)
                Acrescentar Valor (Feature, Classe) aos Resultados
```

Pseudocódigo 1: Método de Seleção de Features

Recorrendo aos três métodos descritos, é possível identificar as *features* mais discriminativas para um dado problema de classificação.

2.3.3. Classificadores

Os classificadores são ferramentas que recorrendo a um número de variáveis, chamadas de *features*, prevêem uma outra variável, a classe. Nos trabalhos de *Barbosa et al* [5], *Go et al* [6], *Pak* e *Paroubek* [7] e outros [8], [16], [14], [17], sobre a análise de sentimento em *tweets,* a utilização de classificadores é apresentada como a melhor solução. A justificação passa por serem considerados dados estatísticos sobre as características linguísticas, permitindo a construção de classificadores independentes da língua e de módulos de NLP. Nos vários trabalhos revistos foram encontrados três classificadores: *Naive-Bayes*[10], *Maximum Entropy*[11] e *Support Vector Machines*[12] (SVM). Este último, devido ao limite de tempo, acabou por não ser considerado para este trabalho, não sendo assim aprofundado nesta secção, mas deverá sê-lo num trabalho futuro.

O classificador *Naive Bayes*, segundo o professor *Dan Jurafsky*[13] da disciplina *From Languages to Information*[14] da Universidade de *Stanford,* é um modelo de classificação simples (“ingénuo”) baseado na regra de *Bayes*. Usualmente conta com uma simples representação (*Bag of Words*) para representar os documentos, e a sua tarefa é prever uma classe onde a escolha se prende com a que obtiver a probabilidade mais alta. Este caso terá as suas desvantagens. Primeiro pode acontecer que a classe com probabilidade mais alta o seja por margem mínima. Em segundo, e mais importante, este classificador não considera a dependência entre as *features*. Na frase “*word’s*

[10] Http://www.stanford.edu/class/cs124/lec/naivebayes.pdf
[11] Http://www.stanford.edu/class/cs124/lec/Maximum_Entropy_Classifiers.pdf
[12] Http://cs229.stanford.edu/notes/cs229-notes3.pdf
[13] Http://www.stanford.edu/~jurafsky/
[14] Http://www.stanford.edu/class/cs124/

best", as palavras serão consideradas como independentes quando aqui existe uma óbvia ligação entre elas.

O funcionamento do classificador traduz-se no cálculo de uma pontuação. Para cada uma das classes consideradas através da equação seguinte, onde *f* representa um array representativo das *features* selecionadas e *c* uma dada classe.

$$\textbf{Pontuação} = argmax_{c \in C}\, P(c) \prod_{f \in features} P(f|c)$$

Relativamente a esta expressão, há dois aspectos importantes. O primeiro diz respeito ao cálculo da probabilidade de uma *feature* dada uma classe, que terá valor diferente de zero se a *feature* existir nessa classe nos dados de treino do classificador, mas e se não existir? Recorrendo a um conjunto de *tweets* para o treino do classificador e a outro para validação dos resultados, existirão *tweets* diferentes em ambos. Poderá não existir uma *feature* nos dados de treino mas sim na validação, dando origem a um problema visto que na fórmula da pontuação, se algum valor no produto for zero o resultado final será sempre zero. Para resolver esta questão existe uma técnica com o nome de *Laplace Smoothing* que adiciona o valor 1 ao numerador e, usualmente, o número total de *features* ao seu denominador. A equação para o cálculo desta probabilidade passa a ser a seguinte:

$$P(f|c) = \frac{\#ocorrência\ da\ feature\ f\ em\ c + 1}{\#total\ palavras\ dos\ tweets\ pertencentes\ a\ c + \#features}$$

O segundo problema prende-se com a multiplicação das probabilidades. Erros do tipo *floating-point underflow* poderão acontecer devido a multiplicações de números muito pequenos. O resultado obtido poderá ser tão pequeno que é transformado no valor zero, o que não é desejado. Para contrariar esta situação é recomendada a substituição das várias multiplicações por somas, e os valores das várias probabilidades pelo seu logaritmo. Esta ideia traduzida na fórmula do cálculo da pontuação apresenta-se na seguinte forma:

$$\textbf{log}(\textbf{Pontuação}) = \log P(c) + \sum_{f \in features} \log P(f|c)$$

O classificador *Maximum Entropy*, ao contrário do classificador anterior, considera dependências entre variáveis. É uma técnica que já deu provas de ser eficiente num

número de aplicações ao processamento de linguagem natural [18]. Foi também anteriormente mostrado [19] que este classificador baseado no Principio da Entropia Máxima [20] obtém, na maior parte das vezes, melhor desempenho na classificação de textos. Sendo F um vector contendo o número de ocorrências das várias *features*, e C uma dada classe, a entropia condicional será calculada através da expressão 1.

$$H(P) \equiv -\sum_{F,C} \tilde{P}(F)P(C|F)\log\big(P(C|F)\big) \quad (1)$$

Este princípio apresenta o problema de encontrar o *P** que maximize a entropia dada por *H(P)*, expressão 2.

$$p^* = arg\max_{p \in C} H(P) \quad (2)$$

O classificador *Maximum Entropy* encontra-se disponível sob a licença *GNU General Public License* pela Universidade de *Stanford*[15] sendo possível utilizá-lo tanto para a o problema da subjetividade, como para o problema da polaridade neste trabalho.

2.5. Avaliação de desempenho

Para justificar os resultados obtidos deverão ser realizados testes, validações, sobre os dados recolhidos. Decidiu-se que os dados recolhidos se deveriam manter separados conforme a sua fonte, e que seria criado um conjunto de dados contendo todos os *tweets* pré-classificados recolhidos, representando todos os dados disponíveis para o projecto. Para a validação dos resultados deverá ser utilizado o método de teste (treino e validação) em que 70% dos dados são usados para treino e os 30% restantes para validação. Esta é a mais forma mais comum de validar um classificador. O classificador será criado recorrendo aos dados para treino, e depois ser-lhe-á aplicado um teste com dados que nunca viu. Este teste será um bom guia sobre o que resultará do classificador quando receber dados novos. Outros métodos poderiam ser utilizados, tais como o *k-fold cross-validation* ou *bootstraping validation*, mas sendo estes mais complexos não existiu tempo para os estudar ficando para trabalho futuro.

Um outro dado prende-se com a execução dos treinos e validações. Estes serão executados sempre sobre três subconjuntos diferentes, tal como se representa na Figura 4.

[15] Http://nlp.stanford.edu/software/classifier.shtml

Figura 2: Modelo sobre a divisão dos dados para as execuções de treino e validação

Nesta Figura 4 estão representadas as divisões sobre os dados que serão feitas para as três execuções com o fim de se apresentarem e analisarem os resultados médios obtidos.

Os resultados finais são naturalmente expressos com contagens ou percentagens sobre as classificações correctas ou erradas. Serão os valores naturais a retirar, mas existem outros a que este estudo recorrerá. Para se afirmar com mais certeza que um dado resultado de classificação é bom ou mau, dois deverão ser utilizados, a sensibilidade e a especificidade [24]. A sensibilidade (ou *True Positive Rate*) mede a proporção de textos bem classificados na classe positiva. A especificidade (ou *True Negative Rate*) mede a proporção de classificações correctas na classe oposta à anterior, a negativa. Formalmente estes dois valores são calculados através das seguintes fórmulas:

$$Sensibilidade = \frac{TP}{TP + FN}, Especificidade = \frac{TN}{TN + FP}$$

Os respectivos valores de TP, TN. FP e FN a quando a classificação sobre as classes X e não X são:

a) Verdadeiros Positivos (TP), prevê-se a classe X e está pré-classificado como X,
b) Verdadeiras Negativos (TN), prevê-se a classe não X e está pré-classificado como não X,
c) Falsos Positivos (FP), prevê-se a classe X mas está pré-classificado como não X,
d) Falsos Negativos (FN), prevê-se a classe não X mas está pré-classificado como X.

Para a visualização destes dados apresenta-se um gráfico com o nome de *Receiver Operating Characteristics Space Graph* [25]. Através dos valores de *True Positive Rate* (TPR ou sensibilidade) e False Positive Rate (FPR ou 1 – especificidade), este

representa o espaço onde se encontram os resultados obtidos. Será assim possível verificar visualmente o compromisso entre a TPR (sensibilidade) e a FPR (1 - especificidade) de um classificador ou de um conjunto de classificadores.

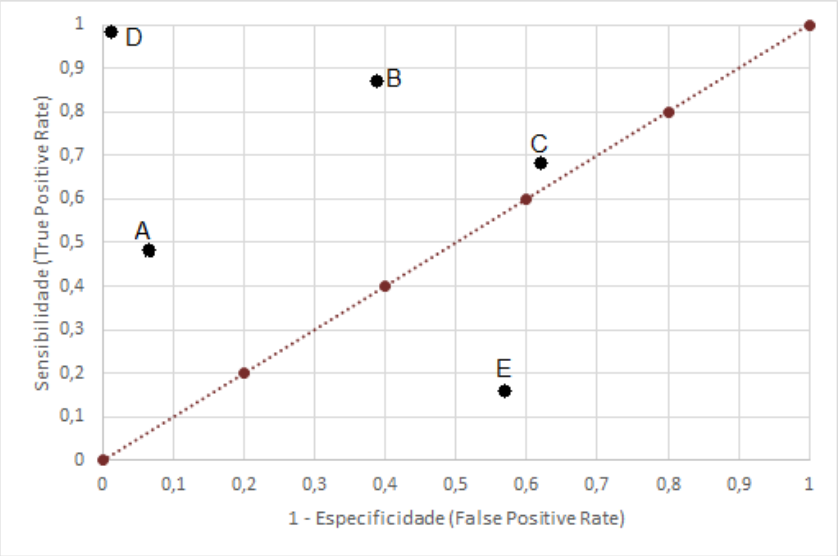


Figura 3: Exemplo de gráfico sobre o espaço ROC

Como exemplo, apresentam-se na Figura 5, cinco classificadores com as marcas de A a E. Um ponto neste gráfico será melhor que outro quanto maior for o valor de TPR e menor for o valor de FPR. Ou seja, temos no ponto D um classificador que apresenta os melhores resultados possíveis. Usualmente, classificadores mais perto do lado esquerdo do gráfico são considerados como mais conservadores. Estes classificam apenas com fortes evidências conseguindo assim poucos falsos positivos. Ao contrário os que apresentam valores mais para o lado direito são considerados mais liberais, podendo apresentar um número maior de verdadeiros positivos, mas também apresentarão um maior número de falsos positivos. Neste exemplo da Figura 5, o classificador A será mais conservador que o B. O classificador C tem resultados que se aproximam da linha x = y representando um que acerta metade das vezes, ou seja que acerta metade das positivas e metade das negativas. Por fim, qualquer classificador que apresente resultados abaixo desta linha representa um classificador pior que um que classifique aleatoriamente. Será de esperar que este segundo triângulo do gráfico seja apresentado sempre sem qualquer classificador.

Um último dado que será verificado tem como nome, curva de aprendizagem. Uma curva de aprendizagem é definida pelo ramo de gestão[16] como uma representação gráfica da aquisição de conhecimento ou experiência ao longo do tempo. Tipicamente

[16] Http://www.qfinance.com/dictionary/learning-curve

na área de *Machine Learning*, uma curva de aprendizagem representa o desempenho geral de um modelo em função do tamanho do corpo de treino [24]. Assim para serem apresentadas e analisadas as curvas de aprendizagem terão de ser executados diversos testes onde se variará a percentagem de dados de treino. O seguinte método poderá ser utilizado para gerar os dados:

I. Para cada tipo de classificação, Subjetividade e Polaridade:
    a. Executar o treino e validação sobre os vários grupos de *tweets*:
        i. Variando a percentagem de dados para o treino do classificador (10, 20, 30, 40, 50, 60, 70%),

II. Determinar todos os dados de validação do classificador:
    a. Percentagens de classificações correctas e erradas
    b. Valores de Sensibilidade e Especificidade

Descreve-se nesta secção todo um processo que deverá ser utilizado para que a validação sobre os classificadores, e os respectivos dados retirados para análise e conclusão, sejam justificados e credíveis.

2.6. Ferramentas disponíveis

Hoje em dia já existem ferramentas que ajudam à realização de cada uma das fases de um processo de análise de sentimento em tweets. Tal como foi revisto em secções passadas, a API do *Twitter* pode ser utilizada para a Recolha de Dados, o *Part-Of-Speech Tagger*[17] pode ser usado no Pré-Processamento, e também o *Stanford Classifier*[18] permite realizar uma Análise Supervisionada.

Ferramentas que realizam todo o processo de análise de sentimento em *tweets* também são encontradas, como por exemplo a apresentada pela empresa DatumBox[19]. Esta, através da sua API, oferece várias opções de análise de textos. Disponibiliza a análise de sentimento em *tweets* tanto para o problema da Polaridade como também da Subjetividade, e ainda a detecção de *spam*, e da língua, entre outras. Um segundo exemplo é a empresa Chatterboxlabs[20] que, não apresentando várias ferramentas como a anterior, se foca na rapidez do seu sistema, que permite processar milhares de textos em segundos, e na qualidade dos seus resultados. Outras empresas como por exemplo a *Semantria*[21] ou a *Salesforce*[22] também fornecem ferramentas para a análise de sentimento em textos.

[17] Http://nlp.stanford.edu/software/tagger.shtml
[18] Http://nlp.stanford.edu/software/classifier.shtml
[19] Http://datumbox.com/features/
[20] Http://chatterbox.co/nlp-and-integration/
[21] Https://semantria.com/
[22] Http://salesforcemarketingcloud.com/

Existem ainda ferramentas disponíveis na internet para a classificação de *tweets* com acesso completamente gratuito. O *Sentiment140*[23], e o *Streamcrab*[24] permitem aos seus utilizadores obter através da introdução de uma palavras-chave, *tweets* classificados quanto à sua polaridade.

Para além das ferramentas até aqui descritas, muitas outras existem[25], cada uma com as suas próprias características, tais como a análise realizada sobre *tweets* passados ou sobre os que são publicados na *live feed* do *Twitter*, ou se permite a visualização geográfica dos *tweets* analisados. O utilizador final tem bastante variedade e deverá verificar qual a que realmente lhe é favorável. No caso da empresa BroadScope, onde foi realizado este estágio, esta prefere desenvolver a sua própria ferramenta visto que considera de importância estratégica a sua conceção e desenvolvimento internos.

2.7. Apreciação Crítica

Ao longo deste capítulo verificou-se que existem vários problemas e várias soluções para a análise de sentimento de *tweets*.

Um caso a resolver passa pela recolha de uma grande quantidade de *tweets* pré-classificados representativos das várias classes, positiva, negativa e neutra. É uma tarefa difícil e que depende das várias suposições que se fizerem sobre estes. Decisões devem ser tomadas para que se utilize uma ou outra técnica, mas principalmente na recolha de *tweets* neutros deverá existir uma maior preocupação e atenção para que os dados representativos desta classe sejam bons. Um outro problema, que se verifica em ambas as abordagens, supervisionada e não supervisionada, prende-se com a reunião de palavras pré-classificadas. Mais uma vez, a pré-classificação, neste caso de palavras, é um dos problemas a resolver, situando-se mais no domínio da linguística, mas também poderá passar pela psicologia. A atribuição de sentimentos a palavras e o uso que os humanos fazem destas é um estudo complexo que não pertence à área da engenharia.

Focando nas abordagens revistas, supervisionada e não supervisionada, num primeiro passo ambas contêm o mesmo problema a resolver. A detecção de dados relacionados com a polaridade pré-atribuída às palavras e as características dos *tweets* existentes nos textos. Verifica-se que uma das análises existentes no modelo apresentado para a abordagem não supervisionada, a lexical, representa o primeiro método a ser utilizado para ambas as abordagens. Na análise lexical encontra-se o problema da pré-classificação de palavras. Deverão ser encontrados vários grupos representativos de

[23] Http://sentiment140.com/
[24] Http://streamcrab.com/
[25] Http://matei.org/ithink/2012/02/08/a-list-of-twitter-sentiment-analysis-tools/

informação útil para classificação tais como palavras positivas e negativas. Posteriormente as abordagens separam-se, onde a supervisionada foca-se na seleção das melhores *features* com o fim de executar o melhor treino possível para o classificador. Enquanto a abordagem não supervisionada continua a trabalhar sobre as várias análises que podem ser aplicadas, sintática e semântica.

Os métodos supervisionados serão os mais adequados à análise de sentimento. A abordagem não supervisionada é sempre considerada como complexa e trabalhosa. Uma possível solução passará por tornar o problema mais simples realizando a análise apenas em textos bem escritos e focando-se sobre palavras que directamente ou indiretamente indicam o tema sobre o qual os textos foram recolhidos.

Capítulo 3

Trabalho realizado

O trabalho realizado focou-se sobre a abordagem supervisionada, com os seus classificadores mas também no passo da análise lexical da abordagem não supervisionada para efeitos de pré-processamento. Ao longo deste capítulo são descritos os vários trabalhos realizados. Inicialmente apresentam-se os requisitos do trabalho, na secção 3.1, seguido da aquisição e caracterização dos dados, na secção 3.2. Posteriormente, apresenta-se o estudado e implementado sobre a análise lexical, na secção 3.3, e os classificadores na secção seguinte, 3.4.. Por fim na secção 3.5 aprecia-se todo o trabalho apresentado ao longo deste capítulo. Uma visão global sobre o sistema descrito ao longo deste capítulo é apresentada esquematicamente no Anexo U.

3.1. Requisitos

Este trabalho enquadra-se no projecto *SoFly Analytics,* da empresa BroadScope, que se encontra em fase de desenvolvimento. O protótipo a desenvolver seguirá a abordagem supervisionada à análise de sentimento de *tweets,* sendo por esta razão dividido em dois submódulos. O primeiro módulo permite o treino do classificador: recebe os textos com a respectiva pré-classificação através de ficheiros de texto, constrói o modelo do classificador e apresenta os resultados de validação obtidos. Este modelo construído é guardado num ficheiro de dados para que posteriormente possa ser lido e utilizado no segundo submódulo, o classificador propriamente dito. Este recebe os textos para classificação também em ficheiros de texto, classifica-os, apresenta os resultados globais e cria um ficheiro de texto contendo cada um dos textos e respectiva classificação. Esta descrição representa o núcleo do trabalho, tendo a sua construção que respeitar algumas imposições:

a) O ambiente de trabalho será a ferramenta (IDE) Visual Studio,
b) A linguagem de programação será *C Sharp* (C#) da *framework .NET*,
c) Os dados a analisar serão apenas *tweets* em língua inglesa.

3.2. Aquisição e caracterização dos dados

Este primeiro ponto prende-se com a obtenção dos dados quer para o estudo, quer para o sistema final. Apresentar e descrever os dados recolhidos e suas respectivas fontes é

o primeiro passo, para posteriormente os analisar com vista a se concluir aspectos referentes à classificação.

3.2.1. Aquisição de Dados

Para a realização do trabalho serão necessários dados com duas finalidades. Uma será a construção e treino do sistema. Outra será a execução do mesmo, primeiro para efeitos de teste e validação e seguidamente para alimentar o sistema em produção. Num primeiro ponto, os dados recolhidos para a construção do sistema incluem:

a) Acrónimos. Sendo os *tweets* textos curtos o uso de acrónimos maximiza o espaço de escrita. Os autores poderão utilizar "*lol*" para em vez de "*laughing out loud*" ou "*anim8*" querendo escrever "*animate*". Foram reunidos 196 acrónimos a partir do *Internet Slang Dictionary*[26],
b) *Stopwords*. Na execução da análise lexical serão detectadas palavras como "*i*", "*myself*", "*this*" ou "*having*", que não contêm qualquer informação sentimental associada. Por esta razão estas palavras não são desejadas para análise. Para que sejam ignoradas, foi construído um dicionário com a junção de dados de duas fontes:
    i. Um dicionário[27] disponível na página na internet: *Webconfs.com*,
    ii. Um trabalho com o nome *Sentiment Analysis*[28] para a disciplina *S11 - Natural Processing Language* do professor *Jason Baldridge* da Universidade de Texas em *Austin*,
c) Palavras com sentimento Positivo ou Negativo, reunidas através de duas fontes:
    i. O trabalho de *Wilson et al.* [26] que representa um estudo sobre a polaridade atribuída a palavras,
    ii. O trabalho da Universidade do Texas referido no ponto anterior,
d) Palavras com sentimento Neutro, reunidas pelo trabalho de *Wilson et al* [26],
e) *Emoticons*, Dicionário criado através dos descritos na Wikipédia[29].

Em segundo lugar, foram recolhidos dados para que fosse possível a realização de treinos, testes e validações sobre o sistema implementado. O corpo de *tweets* pré-classificados foi conseguido através das seguintes fontes:

i. Pelo projecto *Sentiment140* [6], da Universidade de *Stanford*, onde os *tweets* foram reunidos através do uso da API do *Twitter* e pré-classificados recorrendo à existência de *emoticons,*

[26] Http://www.noslang.com/dictionary/
[27] Http://www.webconfs.com/stop-words.php
[28] Http://nlp-s11.utcompling.com/assignments/sentiment-analysis
[29] Http://en.wikipedia.org/wiki/List_of_emoticons

ii. Pelo projecto *TwitterSentiment*[30] de *Niek J. Sanders* contendo *tweets* classificados manualmente,

iii. Pelo trabalho com o nome *Sentiment Analysis*[31] para a disciplina *S11 - Natural Processing Language* do professor *Jason Baldridge* da Universidade de Texas em *Austin*. Foram reunidos *tweets* pela agregação de dois trabalhos, o *Tweet the debates: understanding community annotation of uncollected sources* [27] e o *Characterizing debate performance via aggregated twitter sentiment* [28].

Analisando as palavras pré-classificadas, que no total são 15.011, verifica-se que 6.382 existem em todas as fontes. Comparando a polaridade pré-atribuída a estas 6.382 palavras obtiveram-se 60 que contêm uma polaridade atribuída diferente consoante a fonte. Visto que o trabalho de *Wilson et al*. contém três classes de polaridade (positiva, neutra, negativa) e que este é um estudo completo sobre esta pré-classificação de palavras decidiu-se carregar primeiro todas as palavras deste trabalho. Apenas depois foram adicionadas as palavras conseguidas pelo trabalho da Universidade do Texas e apenas aquelas que não existiam no dicionário já criado. Através destas fontes foi reunido um número considerável de recursos que foram organizados e utilizados no projecto. Na tabela seguinte são descritos os totais sobre os vários dados reunidos.

	QUANTIDADE
ACRÓNIMOS	194
STOPWORDS	661
PALAVRAS PRÉ-CLASSIFICADAS	9613
- Positivas	3172
- Neutras	570
- Negativas	5871
- Com Forte Subjetividade	5552
- Com Fraca Subjetividade	2649
- Sem Subjetividade (Objetivas)	1412
TWEETS PRÉ-CLASSIFICADOS	19814
- Positivos	8609
- Neutros	2081
- Negativos	9124

Tabela 1: Contagens totais sobre os dados recolhidos

3.2.2. *Syntax* e *Meta-Features*

Para ser possivel a classificação de *tweets*, informação sobre estes tem de ser recolhida. Uma possibilidade para esta informação prende-se com as principais caraterísticas dos *tweets*. Desenvolvidas ao longo do tempo, estas características

[30] Http://www.sananalytics.com/lab/twitter-sentiment/
[31] Http://nlp-s11.utcompling.com/assignments/sentiment-analysis

combatem a falta de espaço para escrita e são consideradas como o ruído a quando de uma análise linguística. Pertencem ao grupo das *syntax-features* e a sua existência foi verificada criando a Tabela 2.

***Tweets* (no total são 19.779)**	**Positivos (43,53%)**		**Neutros (10,52%)**		**Negativos (45,95%)**	
Total de características detectadas	97.936		24.188		110.252	
Total de Acrónimos	531	0,54%	47	0,19%	390	0,35%
Total de *StopWords*	60.699	61,98%	17.149	70,90%	70.543	63,98%
Total de *Emoticons* Positivos	8416	8,59%	87	0,36%	444	0,40%
Total de *Emoticons* Negativos	127	0,13%	53	0,22%	8.952	8,12%
Total de *Uppercase*	1.634	1,67%	790	3,27%	1.841	1,67%
Total de *Retweets*	290	0,30%	311	1,29%	196	0,18%
Total de Hiperligações	650	0,66%	789	3,26%	434	0,39%
Total de Repetições de Letras	605	0,62%	8	0,03%	729	0,66%
Total de *Usernames*	5.224	5,33%	970	4,01%	3.597	3,26%
Total de *Hashtags*	442	0,45%	360	1,50%	307	0,28%
Total de Pontuação	5.218	5,33%	385	1,59%	3.918	3,55%

Tabela 2: Distribuição de Syntax-features nos Tweets

Encontram-se no total 250.492 palavras que estão contidas em alguma destas características, representando 79% de todas as palavras existentes nos dados. Analisando cada uma destas características, apresentam-se as seguintes definições, como também a respectiva justificação para a sua representação ou não de uma classe de *tweets* (valores em destaque marcados com cor na Tabela 2):

a) *Emoticons* (*smilies*) são demonstrações de sentimentos expressos pela junção de caracteres de pontuação, por exemplo ":-)" ou ":D". E*moticons* Positivos e Negativos representam respectivamente as classes de *tweets* positiva e negativa onde se detetaram respectivamente em maior número,

b) Acrónimos, curtas palavras que representam uma expressão. Maximizam o espaço para escrita no *tweet*, por exemplo "lol" é traduzido em "*laughing out loud*". Estão presentes em grande número nos *tweets* reunidos, mas representam apenas os mais usuais, tais como "*lol*" ou "*omg*", e alguns mais, não sendo assim representativos de qualquer classe,

c) *Stopwords*, palavras que não contêm qualquer informação sentimental associada, representam a classe de palavras com o maior número detectado. Isto era previsto já que representam uma grande parte das palavras da escrita na língua inglesa,

d) Hiperligações (URLs), por exemplo “http://abc.com/efg”, encontram-se em maior percentagem nos *tweets* neutros,
e) Letras repetidas nas palavras, por exemplo “*looooove*”, representando um enfase dado à palavra “*love*”. Detetadas em número reduzido, não discriminam uma classe de *tweets*,
f) *Usernames*, palavras com uma arroba no primeiro carater, por exemplo “@abcd”. Chamam a atenção de um outro utilizador do *Twitter* para o *tweet* escrito e encontram-se em todas as classes de *tweets*,
g) *Hashtags*, palavras com um cardinal no primeiro carater. Dirigem o *tweet* a um dado tópico, por exemplo “#Coimbra”. Encontram-se em maior percentagem nos *tweets* neutros,
h) Pontuação, poderá ser utilizada pelos autores para dar ênfase a algo, por exemplo “*Sporting!!!!*”. Encontram-se em maior percentagem em *tweets* subjetivos,
i) *Uppercase*, letras em maiúsculas, servem para evidenciar algo que o autor escreveu. Encontram-se em maior percentagem nos *tweets* neutros,
j) Palavras especiais do *Twitter*, por exemplo “RT” que significa *retweeted,* detectaram-se em maior percentagem nos *tweets* neutros.

Além destes indicadores apresentados, outros poderão fazer parte do processo de classificação, como os que constituem um segundo grupo, as *Meta-features*. As *Meta-features* representam informações que se retiram directamente das palavras. A pré-polaridade das palavras e a quantidade destas é apresentada no Anexo A, sugerindo os possíveis indicadores:

a) Palavras positivas, indicador de positividade,
b) Palavras negativas, indicador de negatividade*,*
c) Palavras subjetivas, palavras que representam algum tipo de sentimento, seja este positivo ou negativo,
d) Palavras objetivas ou palavras pouco subjetivas como se verifica no trabalho de *Wilson e tal.* [26], indicadores de objetividade e portanto sem um sentimento atribuído (neutro).

Dois dados importantes a referir neste ponto serão os números de *tweets* negativos e de palavras negativas. Como foi apresentado na Tabela 1: o número de palavras negativas reunidas é maior que o número das positivas, e também o número de *tweets* negativos reunidos é maior que os positivos. Este é um dado importante visto que deverá posteriormente fazer a diferença na aprendizagem sobre os *tweets* negativos. Esta classe de *tweets*, visto que existem mais dados sobre ela, deverá obter melhores resultados a quando da classificação.

Por outro lado existem outras características relacionadas com as palavras que não serão consideradas como parte das *meta-features,* mas que poderão ser usadas, tais como as classes gramaticais. Dados sobre as várias classes gramaticais, nomes, verbos, adjetivos, poderão ser relevantes para uma boa análise do sentimento do *tweet*.

3.2.2.1. Part-Of-Speech Tagging

Na recolha de informação sobre as palavras de um texto existem dados sobre a classificação gramatical que uma palavra terá no contexto em que foi escrita. Existem várias classes gramaticais de palavras, como nomes, verbos, adjectivos, mas quais serão as mais importantes para a análise sentimental e como se obtêm? Tal como em estudos revistos ([7]), um dado a analisar para a classificação de sentimento em *tweets* é a utilização do processo de *POS-Tagging*. O *Part-Of-Speech Tagging* é definido como um processo que atribui a palavras as respectivas classes gramaticais. Nasceu através do desejo de se colocar um computador a entender a língua humana e pode ser obtido através da Universidade de *Stanford*. O grupo do processamento da linguagem natural desta universidade já estudou e implementou vários processos sobre o tema da linguagem natural e um deles é este método, o *POS-Tag*[32]. Para uma visualização ao pormenor de todas as classes de palavras e os seus respectivos diminutivos apresenta-se a tabela do Anexo C.

Um estudo sobre a construção e funcionamento interno deste processo não está no âmbito deste trabalho, mas a sua utilidade para a análise de sentimento em *tweets* está. Pretende-se verificar quais as classes gramaticais representativas de *tweets* positivos, neutros e negativos. Com a divisão dos *tweets* em dois grupos, o dos *tweets* subjetivos e o dos objetivos, verificam-se quais as classes gramaticais que melhor os representam, informação essa que se encontra descriminada no Anexo D. Separa-se novamente os *tweets* em outros dois grupos, o positivo e o negativo e repete-se o processo, obtendo a informação gramatical descriminada no Anexo E. Através destes dois estudos das classes gramaticais presentes nos vários grupos de *tweets,* concluiu-se quais as mais representativas quanto à classificação subjetividade e à polaridade. Estes dados foram calculados utilizando a fórmula seguinte:

$$P^{Tag}_{Classe1,Classe2} = \frac{P^{Tag}_{Classe1} - P^{Tag}_{Classe2}}{P^{Tag}_{Classe1} + P^{Tag}_{Classe2}}$$

[32] Http://nlp.stanford.edu/software/tagger.shtml

Onde os valores de $P^{Tag}_{Classe1}$ e $P^{Tag}_{Classe2}$ representam respectivamente a percentagem da ocorrência de uma dada *Tag* gramatical em uma das duas classes de *tweets*.

No caso do Anexo D, o valor 1 representa o grupo de *tweets* subjetivos e -1 o grupo dos objetivos. Em ambos se verifica a existência de algumas classes de palavras que serão as mais representativas de uma dada classe de *tweets*. Ao detalhe, verifica-se que a classe de textos subjetivos se deverá representar por:

a) Nomes que terminam com a letras "s" (POS), chamadas de terminações possessivas como "*José's car*",
b) Exclamações (UH), afirmações,
c) Estrangeirismos (FW). Poderão ser confundidas com erros ortográficos,
d) Advérbios (RB), usualmente modificam um verbo, como os de intensidade "mais" ou "menos",
e) Numerais Ordinais (LS), tais como "primeiro", "segundo", "terceiro",
f) Adjectivos (JJ) ou adjectivos comparativos (JJR), geralmente para atribuir propriedades a palavras,
g) Verbo (PRP) na primeira pessoa ou na segunda para um autor de um *tweet* se dirigir ao seu público.

Já para o caso dos *tweets* objectivos, estes serão formados maioritariamente com os seguintes indicadores:

a) Pronome com "*wh*" possessivo (WP$), por exemplo "*whose*" que representa a forma possessiva de "*who*",
b) Adverbio, superlativo (RBS), por exemplo com o uso de "most" ou "least",
c) Nomes próprios no plural (NNPS), nomes como "*Americans*", e no singular (NNP),
d) Numeral cardinal (CD), "*ten*", "*million*" ou "*forty-two*" ou até "*mid-1980*",
e) Adjectivo superlativo (JJS), expressão geralmente usada para exprimir opiniões,
f) Pronomes com "*wh*" (WP), como "*whatsoever*", "*who*", "*what*".

Em suma, os *tweets* objectivos (neutros) serão um grupo de *tweets* que apresentam maioritariamente pronomes, nomes próprios, advérbios superlativos tal como adjectivos e palavras que contêm alguma forma numérica. No caso dos subjetivos, serão aqueles que mais provavelmente contêm exclamações, afirmações, contêm nomes terminados com a letra "s" e muitos adjectivos na sua forma normal e comparativa, tal como pronomes que não estejam na terceira pessoa.

No gráfico no Anexo E, o valor 1 representa os *tweets* positivos e o -1 os negativos. Comparando os grupos de *tweets* presentes no Gráfico 2, os indicadores para o grupo dos positivos:

a) Nomes próprios no plural (NNPS) ou singular (NNP) ou até nomes que terminam com a letras “s” (POS),
b) Adjectivo superlativo (JJS), por exemplo “*calmest*”, “cheapest”, “*choicest*”,
c) Exclamações (UH), afirmações,
d) Adverbio comparativo (RBR), tal como “*greater*” ou “*heavier*”,

No caso do grupo dos *tweets* negativos, este conterá:

a) Numerais Ordinais (LS), tais como primeiro, segundo, terceiro,
b) Verbo no passado (VBD) ou no particípio passado (VBN),
c) Advérbios (RB), usualmente modificam os verbos como por exemplo os de intensidade: “mais”, “menos”
d) Chamados de “*Particle”* (RP), palavras curtas, muitas vezes ligadas a verbos, como por exemplo “*aboard*”, “*into*”, “*down*”, “in”, “off”, “on”, “out”,
e) Estrangeirismos (FW).

Estes indicadores podem ser comprimidos no caso dos *tweets* positivos em nomes próprios, adjectivos e exclamações e para os negativos, em forma verbais no passado, marcadores de listas ou advérbios.

Comparando este estudo com os resultados encontrados no trabalho de *Alexander et al.* [7] e com o trabalho de *Spencer et al.* [29], verifica-se que existem diferenças na análise dos *tweets.* Este dado poderá ser previsível visto que os *tweets* são diferentes em cada estudo. Mesmo assim, existe muita informação igual o que também seria previsível sendo que todos os trabalhos são referentes à língua inglesa.

A maior semelhança é referente aos *tweets* positivos e negativos. Para a classe dos positivos, as principais classes gramaticais discriminativas são nomes, adjetivos e advérbios. Estas também se encontram presentes nos vários trabalhos existindo, para o caso dos adjetivos, diferenças na sua forma. Para o caso da classe negativa, o mesmo acontece com os verbos na forma passada e particípio passado, com os advérbios e as palavras curtas ligadas aos verbos (as *Particles*). Estas classes de palavras foram verificadas como as discriminativas da classe de *tweets* negativos tal como se verifica neste trabalho.

Para o caso dos *tweets* subjetivos e objetivos, existiu uma diferença maior. Ao contrário das classes de *tweets* anteriores, para estas existiram apenas três classes gramaticais presentes em todos os trabalhos: pronomes possessivos, advérbios e interjeições. Neste caso, mais foram encontradas, mas estas foram variando de trabalho para trabalho.

Apesar das diferenças atinge-se uma conclusão positiva. A informação gramatical presente nos *tweets* poderá ser útil para a sua classificação. O uso de *POS-Tagging*

sobre as palavras poderá ser uma mais-valia sendo importante verificar os resultados que se obterão com estas informações.

3.2.3. *N-Grams*

Outra possível abordagem para a separação das classes de *tweets* prende-se com as próprias palavras, e as sequências por estas formadas, com o nome de *N-grams*[33]. Estes nasceram a partir de modelos estatísticos da língua que recorrem a cadeias de *Markov*. Calcular a distribuição estatística de sequências de palavras sobre uma dada língua poderá oferecer a possibilidade de, por exemplo, prever qual a palavra seguinte numa frase. Neste trabalho a intenção não é verificar as probabilidades das possíveis sequências de palavras, mas sim utilizar estes dados como possíveis indicadores de uma classe de *tweets* recorrendo a:

a) *Unigrams*, uma só palavra,
b) *Bigrams*, sequências de duas palavras,
c) *Trigrams*, sequências de três palavras.

Será possível, recorrendo a estas sequências de palavras, retirar mais dados sobre os *tweets* para a sua classificação? No trabalho de *Alexander et al.* [7] é apresentado que recorrendo a *n-grams* de ordem elevada, como os *trigrams,* se capturam padrões que representarão expressões atribuídas a sentimentos positivos ou negativos. Por outro lado, os *unigrams* serão os que providenciarão um maior detalhe sobre os dados e os *bigrams* deverão ser importantes para a captação da negação. É referido que recorrer a *N-grams* aumentará a precisão da classificação, devendo-se assim analisar as várias possibilidades para o treino dos classificadores. Obter-se-ão melhores resultados apenas com o uso dos *unigrams*, ou será melhor utilizar estes juntamente com o *bigrams*, ou até os *trigrams?* Existem várias perguntas que poderão ser respondidas aquando do uso destes *n-grams*, e algumas das principais combinações são as presentes na Tabela 3, onde A, B e C representam palavras não pertencentes ao grupo das *Stopwords*, ou seja como é visível na Tabela 3, existe um vocabulário para os *unigrams* constituído por todas as palavras diferentes existentes nos textos recolhidos e mais uma, a "*stopword*", representando essas. Além disto, os dados desta tabela são referentes a todos os *tweets* reunidos para este trabalho.

[33] http://link.springer.com/chapter/10.1007/978-1-4471-6308-4_6#page-1

Grupo de *N-Gram*	**Exemplo**	**Tamanho Vocabulário**
Unigrams,	"A", "B", "C", "*Stopword*"	25199
Unigrams – StopWords,	"A", "B", "C"	25198
Unigrams + POS-Tagging,	"A_Tag", "*Stopword_Tag*"	31765
Unigrams + POS-Tagging – StopWords,	"A_*Tag*", "B_*Tag*"	31602
Bigrams	"A B", "*Stopword* A"	72929
Bigrams – StopWords,	"A B"	47080
Trigrams,	"A B C, "*Stopword* B *Stopword*"	118346
Trigrams – StopWords.	"A B C"	0

Tabela 3: N-grams

Sabendo que as *Stopwords* não contêm informação sentimental, os vários conjuntos que as contêm não serão considerados. Grupos contendo este tipo de palavras conterão informação que deverá ser desnecessária para o tipo de classificação que se pretende, além de que mais dados para processamento se traduz em mais tempo gasto.

3.2.4. Outras *Features*

Sendo a polaridade pré-atribuída das palavras um dos indicadores de classe de *tweets*, um factor importante será a negação. Para se exprimir algo positivo, ou se recorre directamente a palavras positivas, ou nega-se as negativas, como acontece também para o caso oposto. *Richard Nordquist*[34] afirma que existem dois tipos de negação de frases na lingua inglesa:

a) Negação com "*not*" e "*n't*"
b) Negação com palavras: "*never*", "*neither*", "*nobody*", "*no*", "*none*", "*nor*", "*nothing*", "*nowhere*", "*seldom*" e "*hardly*".

Verifica-se que estas palavras não tendo uma polaridade pré-atribuída estão contidas no dicionário das *Stopwords*. Retirando-as deste dicionário, cria-se um dicionário de negações. Analisando os *tweets* quanto a estas negações, obtiveram-se os dados apresentados na tabela do Anexo B.

Se na Tabela 2, se verifica que os *tweets* neutros são os que contêm menos *Stopwords*, este dado traduz-se no Anexo B em menos negações detectadas. Os *tweets* da classe neutra existem em menor número, mas tal como se verifica na Tabela 2, e são os que contêm mais Uppercases, Hashtags, Hiperligações. Não fazendo sentido negar estes termos, é retirado espaço para a escrita, obtendo-se poucas negações positivas. Existindo mais *tweets,* caso dos positivos, detetam-se mais negações, tal como nos *tweets* negativos, onde o número mais que duplicou em relação às negações

[34] Http://grammar.about.com/od/rs/g/Sentence-Negation.htm

detectadas nos positivos, de onde se poderá retirar a seguinte hipótese: as negações são utilizadas principalmente para expressar algo negativo, enquanto se recorre a palavras com sentido positivo para expressar algo positivo.

As principais palavras utilizadas nos *tweets* disponiveis foram as com "*n't*" e as palavras "*not*" e "*no*". Este dado seria previsível visto que estas serão as principais negações. Para comprovar estes dados verificou-se quais as principais palavras escritas no *twitter* segundo *Lev Grossman*[35], e os resultados estão de acordo. A principal negação é realizada através do uso de "*n't*", seguido da palavra "*not*" e por fim "*no*". Poder-se-á resumir que as negações, principalmente as três palavras referidas, serão bons indicadores do uso das referidas negações, devendo ajudar na classificação.

Sobre a negação existe ainda a possibilidade do uso da segunda negação. Será que o uso de várias negações na mesma frase existe? Será um dado relevante a ter em conta? Analisando os *tweets* de que se dispõe, verifica-se que existem frases nos *tweets* que contêm dupla negação. Em números, este dado traduz-se em que nos 8609 *tweets* positivos existem 37 frases com tal característica enquanto nos 2081 neutros existem 12 e nos 9124 negativos 95, o que são números baixos. Mesmo considerando que um *tweet* será apenas uma frase, o que poderá nem sempre acontecer, a percentagem de *tweets* com esta dupla negação é baixa.

Um outro dado que poderia marcar a diferença é o tópico a que os textos se referem. Se este for conhecido à partida, talvez focando a análise sobre este se obtenham resultados mais precisos, como foi considerado por *Jiang et al.* [10]. Existindo um tópico definido para cerca de 68,8% dos *tweets* disponiveis*,* foi possível analisar directamente a existência deste nos *tweets,* mas os resultados foram baixos. Apenas 2,2% dos *tweets* contêm o seu tópico escrito directamente no texto. Este resultado comprova o que *Jiang et al*. apresenta para a classificação de *tweets* através de um *target* (tópico). Será necessário usar não apenas o tópico, mas também os seus subtópicos. Será natural o autor de um *tweet* expressar algum sentimento sobre um *target* comentando não sobre este, mas sim sobre algo relacionado. Por exemplo, poderá ser expresso um sentimento em relação a uma empresa através dos seus produtos ou tecnologias. Para realizar uma análise sobre este tema, será necessário saber previamente os subtópicos, o que poderá nem sempre acontecer. Ou, como previsto para este trabalho, a própria escolha dos *tweets* já subentende o tópico a que se referem. Por esta razão, o estudo da análise sentimental de *tweets* baseada em *targets* tem baixa prioridade, o que não quer dizer que não possa vir a ser realizada no futuro.

[35] http://techland.time.com/2009/06/08/the-500-most-frequently-used-words-on-twitter/

3.3. Pré-Processamento

A análise lexical representa uma primeira abordagem simples à tentativa de retirar um sentimento de textos. Aplicando alguns métodos de NLP, recolhem-se dados sobre as palavras existentes no texto com o fim de calcular o valor a atribuir ao sentimento positivo, neutro e negativo. Usualmente escolhe-se o de maior valor, atribuindo assim um sentimento ao texto. Para se visualizar o processo que será descrito a seguir apresenta-se a Figura 4.

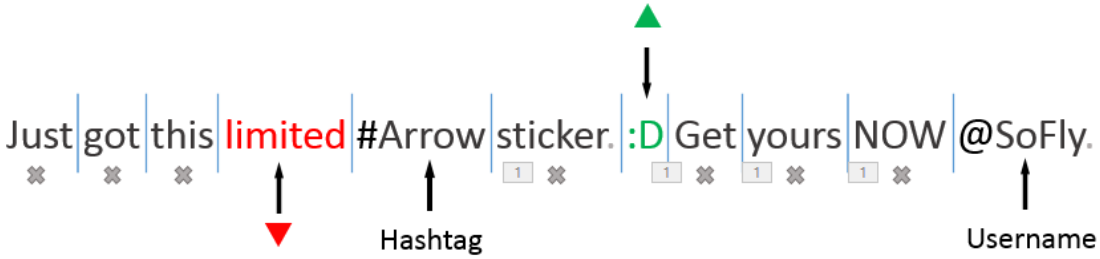


Figura 4: Análise Lexical a um Tweet

O primeiro passo consiste na detecção e remoção dos *emoticons* existentes no texto, que são detetados e removidos em primeiro lugar devido ao seu formato. Verifica-se directamente através destes se o autor do *tweet* expressou algum tipo de sentimento e se foi positivo ou negativo. O passo seguinte passa consiste em retirar os restantes sinais de pontuação. Na Figura 4: Análise Lexical a um Tweet esta pontuação é representada pelo caractere ponto final.

Sem estes caracteres nos textos é possível aplicar o método chamado de *Tokenization*[36] que tem como objectivo partir os textos pelas suas várias palavras ou *tokens,* resultando numa lista contendo todas as palavras para que seja realizada a procura de pré-polaridade. Esta procura é altamente dependente da quantidade de palavras com pré-polaridade atribuída que estão disponíveis. Assim, existirá sempre a possibilidade de palavras não serem detectadas por não constarem nos dicionários. Os dicionários conterão muitas palavras, mas muitas mais existirão na língua inglesa. Para maximizar o número de palavras encontradas, e se conseguir obter melhores resultados, dois processos poderão ser incluídos, o *Stemming* e a *Lemmatization*, descritos na secção seguinte, 3.3.1.. Na procura por palavras com polaridade atribuída encontrar-se-á no exemplo da Figura 4 apenas uma, a "*limited*", que contém uma polaridade pré-atribuída pelos dicionário disponível como negativa.

[36] http://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

Por fim, obtendo as contagens de palavras com polaridade, como também a contagem sobre os *emoticons*, é calculado um valor relativo a cada uma das classes. Verificando qual o maior valor, atribui-se a sua respectiva classe ao texto. Este processo apresenta-se como o mais simples para a extração de sentimento de textos, trazendo com ele o pressuposto de que um texto com mais palavras positivas do que negativas será um texto positivo e vice-versa. Torna-se assim um processo falível visto que um texto, apenas porque contém palavras com uma polaridade atribuída de positiva, não quer necessariamente dizer que seja um texto positivo. Existem vários aspectos mas complexos que devem ser verificados, mas estes prendem-se com a análise seguinte, a sintática.

3.3.1. Stemming e Lemmatization

Para tentar maximizar o número de palavras encontradas, e se conseguir obter os melhores resultados, dois processos podem ser incluídos, o *Stemming* e a *Lemmatization*.

O *Stemming* (radicalização) corresponde ao processo de remover as terminações morfológicas e verbais mais comuns das palavras para se encontrar o *stem* (o radical) da palavra. O algoritmo mais popular para o cálculo dos radicais na língua inglesa é o de *Porter* [30]. Este algoritmo opera em vários passos, definindo regras baseadas nos sufixos existentes na língua inglesa, e que são aplicadas às palavras. Com este processo, ao ser detectada a palavra “liked” é removido o sufixo “ed”. O seu radical “like” é igual à palavra “like”, atribuindo assim o mesmo sentimento. Palavras como “liked”, “likely”, “likeness” ou até “liking” resultarão com este processo na palavra “like”, levando à atribuição da mesma polaridade a todas elas. Estes são exemplos de casos ideais para este processo, mas existirão casos em que o radical da palavra não será igual à palavra base. Sendo constituído apenas por alguns carateres iniciais da palavra, o que pode resultar na deteção de palavras por um dado radical que não pertencem à mesma família. Apesar desta limitação, a utilização deste processo deverá ser útil, visto que resultará num aumento do número de palavras detectadas. Aplicando este processo sobre a deteção de palavras obtiveram-se os dados apresentados na tabela no Anexo F. Existirem palavras que não foram detectadas é um resultado esperado. Os ficheiros de validação do processo de *Stemming* contêm cerca de 23.000 palavras e respectivos radicais, mas a língua inglesa terá muitas mais. Para o caso em que os radicais são encontrados mas as respectivas palavras não, definiu-se que aplicar um processo de verificação de similaridade entre palavras. Definindo que uma palavra é similar a outra se obtiver um valor de similaridade maior que 95%, obtiveram-se os dados da tabela no

Anexo G. No total, aplicando este processo, o valor de palavras não encontradas desceu para todas as três classes de *tweets*:

a) Positivas com ganho ≈ 12%,
b) Neutras com ganho ≈ 13%,
c) Negativas com ganho ≈ 11%.

Não são valores muito altos mas conseguiram-se encontrar mais palavras do que as disponíveis inicialmente, o que confirma o valor deste processo, que será uma mais-valia para a detecção de palavras.

O segundo processo tem como nome *Lemmatization*[37]. Este corresponde a uma evolução do processo anterior e tem como objetivo encontrar o *lemma* de uma palavra, ou seja, a sua forma base, como se encontra nos dicionários de qualquer língua. Utilizando uma análise vocabular e morfológica, e recorrendo a algoritmos de aprendizagem, consegue-se devolver a base de qualquer palavra recebida resultando na eliminação dos erros que o processo de *Stemming* apresenta.

Um exemplo para se verificar que este processo será melhor poderá ser a palavra "*saw*". O processo de *Stemming* recebendo esta palavra retornará "saw", não sendo de qualquer utilidade na procura de palavras. Utilizando *Lemmatization* obter-se-á a sua forma base, "*see*", sendo assim possível verificar se existe alguma polaridade atribuída.

Tendo como linguagem de programação para o projecto a linguagem C#, encontra-se disponível uma biblioteca com o nome de *LemmaGen*[38] que deverá ser incluída. Esta biblioteca contém um algoritmo de aprendizagem que foi treinado para gerar as regras que uma dada língua necessita para o seu processo de *Lemmatization* [31]. É referido que a sua fonte de dados para o caso da lingua inglesa contém cerca de 70.000 formas de palavras e 27.000 *lemmas,* conseguindo com estes uma validação com precisão de 99% na obtenção de *lemmas.* Este resultado é muitíssimo bom. Realizando novamente os testes com os resultados apresentados no Anexo A, e acrescentando os resultados obtidos com os dois processos descritos nesta secção gerou-se a tabela no Anexo H.

Os resultados demonstram que o processo de *Lemmatization* aumentou o número de palavras detectadas tendo com este os seguintes ganhos:

a) Positivas com ganho ≈ 18%,
b) Neutras com ganho ≈ 22%,
c) Negativas com ganho ≈ 12%.

Com o uso de *Lemmatization* em média foi obtido um ganho na deteção de palavras com polaridade em cerca de 18,05%, contra o *Stemming* que resultou numa média de 12%. Ambos são resultados positivos visto que resultam num aumento do número de

[37] Http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
[38] Http://lemmatise.ijs.si/

palavras detectadas e ambos podem ser utilizados, mas confirma-se que a *Lemmatization* será a mais favorável.

3.4. Seleção de *Features*

Para se efectuar uma análise às diferentes *features* existentes e selecionar as mais discriminativas, foram utilizados os métodos da informação mútua e do Chi-Square apresentados no capítulo anterior.

A Tabela 4 apresenta o *Top* 10 das *features* mais discriminativas (com valor mais alto), para a classificação dos *tweets* quanto à subjetividade e à polaridade. As *features* são apresentadas por ordem decrescente dos valores correspondentes.

Chi Square		**Informação Mútua**	
Classificação Subjetividade	**Classificação Polaridade**	**Classificação Subjetividade**	**Classificação Polaridade**
URLs	*Emoticons* Positivos	*URLs*	*Emoticons* Positivos
Retweets	*Emoticons* Negativos	*POS-Tag: IN*	*Emoticons* Negativos
POS-Tag: IN	Palavras Negativas	*Retweets*	Palavras Negativas
POS-Tag: NNP	Negações	*POS-Tag: NNP*	Negações
Hashtags	Palavras Negadas	*Hashtags*	Pontuação
POS-Tag: RB	Pontuação	*POS-Tag:* RB	Palavras Negadas
POS-Tag: NNPS	Palavras Positivas	*Uppercases*	Palavras Positivas
Uppercases	*POS-Tag:* RB	Palavras Subjetivas	*POS-Tag:* RB
Palavras subjetivas	*POS-Tag:* VBD	*POS-Tag: NNPS*	*POS-Tag:* VBD
Palavras Objetivas	*POS-Tag:* FW	Palavras Objetivas	*POS-Tag:* FW

Tabela 4: Top 10 sobre Seleção Features

Comparando os resultados apresentados neste *Top* 10, verifica-se que várias *features* foram selecionadas por ambos os métodos em cada um dos problemas, nomeadamente as *hashtags, retweets, Urls, Uppercases*, e as *features* de *POS-Tag* no caso da subjetividade e as palavras pré-classificadas (positivas e negativas), os *emoticons* (positivos e negativos), negações, palavras negadas e *POS-Tags* no caso da polaridade.

Outro método de seleção de *features* é baseado no cálculo da frequência de cada uma. Representa a abordagem de seleção mais simples e foi apresentado na secção 2.3.2. Uma outra possibilidade passa pela realização de uma procura exaustiva, registando os resultados obtidos por um classificador recorrendo às várias combinações de *features* em cada um dos problemas, subjetividade e polaridade. Este caso levanta o problema do tempo de execução. Existindo trinta *features* possíveis para cada problema,

discriminadas no Anexo I, o número de combinações a experimentar é enorme e consequentemente o tempo de execução da procura também o é.

$$\#Combinações = 2^{30} - 1$$

A solução para se conseguir realizar esta procura em tempo razoável passa pela redução do número de combinações. Ou seja, agrupar-se estas *features* em vários grupos de *features* semelhantes, reduzindo assim o número de combinações a experimentar. Por exemplo, a junção das *features*: palavras positivas e negativas, criará o grupo das palavras com polaridade pré-atribuída. No problema da subjetividade, estes grupos traduzem-se em catorze, e para a polaridade em quinze.

Tendo reduzindo o número de combinações já foi possível executar, em tempo aceitável, uma procura exaustiva pelas melhores *features* tendo obtido os seguintes resultados para o problema da subjetividade:

-Grupo 2: *Hashtags*, *Retweets*, *Uppercase* e *URLs*,

-Grupo 7: *POS-Tag* LS (*List item marker*) e JJS (*Adjective superlative*),

-Grupo 9: *POS-Tag* PRP (*Personal pronoun*) e NNP (*Proper Noun, singular*),

-Grupo 10: *POS-Tag* VBP (*Verb, non-3rd person, singular present*) e IN (*Preposition*).

Estes dados estão de acordo com os da Tabela 2 que foram apresentados no ponto sobre a aquisição e caracterização dos dados. As *features* do grupo 2 seriam as que descriminariam melhor os *tweets* objectivos ajudando na classificação da subjetividade. Tal como as várias *POS-Tags* que foram escolhidas. Apesar de não serem as que se verificaram em maior número, estão de acordo com o gráfico no Anexo D, mostrando que este tipo de *features* gramaticais é importante para uma boa classificação.

Realizada a mesma procura pelas melhores *features* para o problema da polaridade dos *tweets* obtiveram-se as seguintes*:*

-Grupo 1: Palavras com polaridade (positivas, negativas),

-Grupo 2: Emoticons com polaridade (positivos, negativos),

-Grupo 9: *POS-Tag* RBR (*Adverb, comparative*) e FW (*Foreign word*),

-Grupo 10: *POS-Tag* PDT (*predeterminer*) e TO (*word "to"*),

-Grupo 11: *POS-Tag* WPS (*possessive "wh"-pronoun*) e VBZ (verb*, 3rd person, singular present*)

Tal como foi verificado e apresentado na Tabela 2, as *features* do grupo 2 são as que descriminam melhor os *tweets* positivos e negativos. Comprovaram-se também os dados no Anexo A, visto que o grupo 1, as palavras pré-classificadas, também

pertencem às melhores *features* para a classificação da polaridade, tal como as várias *POS-Tags* que foram escolhidas, que estão de acordo com o gráfico do Anexo E.

3.5 Classificadores

A aplicação do classificador *Naive Bayes,* como de qualquer outro classificar supervisionado, passa por dois processos representantes da criação do modelo (treino) do classificador e, posteriormente, o uso do próprio classificador, Figuras 5 e 6.

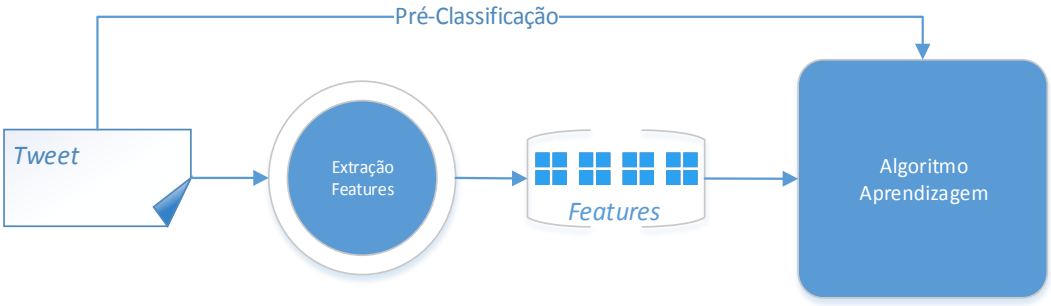


Figura 5: Modelo de Aprendizagem de um Classificador

A Figura 5 mostra que para cada *tweet* existirá um pré-processamento com o fim de se retirarem todas os dados sobre este, as suas *features*. Juntamente com a sua pré-classificação, o algoritmo de aprendizagem criará um modelo de classificação que posteriormente permitirá classificar novos dados, seguindo o modelo apresentado na Figura 6.

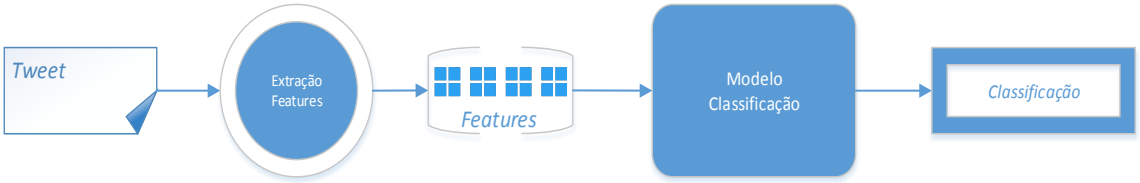


Figura 6: Modelo de um Classificador

Após o pré-processamento de cada *tweet* de onde se retiram as *features* deste, realiza-se a classificação. Os dois classificadores utilizados foram o *Naive Bayes* e o *Maximum Entropy*. A principal diferença entre os dois reflete-se no modo como olham para as *features* que recebem. No primeiro caso, o *Naive Bayes,* as *features* são tratadas como independentes umas das outras, enquanto o classificador *Maximum Entropy* considera dependência entre *features*. Apesar do seu pressuposto simples, o *Naive Bayes,* poderá ser um bom algoritmo para a resolução do problema da classificação dos *tweets*. sendo um algoritmo rápido e não tendo um alto custo de memória. Através da regra de *Laplace*

Smoothing tornar-se robusto contra *features* que não tenha encontrado antes. Para criar um classificador de *Naive Bayes* são necessários dois algoritmos como já referido. Primeiro um para o treino e em segundo para o classificador em si. Para o primeiro caso, o algoritmo implementado corresponde ao Pseudocódigo 4, o segundo ao Pseudocódigo 5.

```
Algoritmo de Treino (Tweets de Treino, Features, Classes de Tweets):
1. Para cada Tweet de Treino:
        1.1 Contar o número de ocorrência de cada Feature
        1.2 Adicionar estes números às contagens globais
2. Para cada Classe:
        2.1 Calcular a probabilidade (Prior Probability) correspondente
        2.2. Para cada Feature:
                2.2.1 Calcular a Probabilidade Condicional entre Feature e Classe
```

Pseudocódigo 4: Método Treino Classificador Naive Bayes

```
Algoritmo de Classificação (Tweets de Teste, Classes de Tweets):
1. Para cada Tweet de Teste:
        1.1 Para cada classe:
                1.1.1 Calcular as Probabilidades Condicionadas
                1.1.2 Calcular o Score
2. Verificar qual o maior Score
3. Atribuir ao Tweet a Classe correspondente ao maior Score
```

Pseudocódigo 5: Classificador Naive Bayes

Relativamente ao classificador *Maximum Entropy,* foi utilizada a implementação disponibilizada pela Universidade de *Stanford*[39] sob a licença *GNU General Public License*.

3.6 Apreciação Crítica

Tendo apresento todos os dados de relevo sobre o trabalho realizado, é feito nesta secção um pequeno resumo conclusivo sobre todo o capítulo.

Na recolha e caracterização dos dados comprovam-se que é possível obter muita informação a partir de *tweets* com vista à análise de sentimento e respectiva classificação, quer recorrendo a indicadores sobre as várias características dos próprios *tweets*, como as *hashtags*, *emoticons* ou os acrónimos, quer a partir de características gramaticais dos textos, como a polaridade pré-existente das palavras, quer ainda através da verificação de sequências de palavras, os *N-Grams*.

[39] Http://nlp.stanford.edu/software/classifier.shtml

Recorrendo a métodos de *Feature Selection* verificou-se que as *syntax-features* constituem um bom grupo de *features* a serem utilizadas qualquer que seja a representação do *tweet, N-grams* ou *arrays* contendo informação binária ou real. Ambos recorrem a este grupo de *features* para bem separar as classes que se deseja aprender.
A recolha de dados pré-classificados, quer dos *tweets*, quer das palavras, é um aspecto que requer muita atenção. Em ambos os casos existe muita subjetividade, sendo que, por exemplo, duas pessoas olhando para o mesmo texto ou a mesma palavra poderão apresentar classificações diferentes. Para potenciar a detecção de palavras, abordaram-se os processos de *Stemming* e *Lemmatization*.
Seguindo por uma abordagem supervisionada, há que ter em atenção o tempo necessário para treino, dependendo da quantidade de dados e do número de *features,* será maior ou menor. Além disso, o tempo de treino depende ainda dos classificadores, enquanto o *Naive Bayes* não considera dependências entre *features,* o aumento da complexidade associado ao algoritmo do classificador *Maximum Entropy* deverá traduzir-se por melhores resultados, mas também por tempos de aprendizagem mais elevados. Além disto, as abordagens supervisionadas, recorrendo a dados concretos para treino poderão posteriormente obter bons resultados apenas para dados semelhantes aos do treino. Seja, por exemplo, em relação ao estilo de escrita, seja em relação ao contexto apresentado nos textos. Apesar de tudo, será sempre um problema de classificação e não linguístico, o que é desejado, simplificando o problema e dando origem a melhores resultados.
No caso da abordagem não supervisionada, se por um lado não engloba estes problemas relativos a dados e treino, representa por outros problemas mais complexos a resolver. Os problemas associados à análise lexical são relativamente simples. Prendem-se principalmente com a recolha de dados pré-classificados e a sua detecção, mas para a análise sintática, apesar de nos dias de hoje existir muita tecnologia que pode ajudar a recolher os dados gramaticais sobre os textos, estes têm de ser tratados e analisados, representando este aspecto algo que precisa de ser estudado com grande detalhe. Na língua inglesa, imagine-se que apenas se utiliza quatro classes de palavras: nomes, verbos, adjetivos e pronomes. Terá de existir um estudo linguístico com o intuito de perceber o que se poderá saber a partir do uso destas classes de palavras e como retirar informação relevante para a classificação. Isto representa apenas uma pequena percentagem de trabalho. O número de classes gramaticais é muito maior traduzindo o problema num muito maior e mais complexo, visto que também as próprias ligações entre palavras são importantes. Para se utilizar esta abordagem na classificação dos *tweets*, muitos aspectos sobre a língua teriam de ser estudados, algo que não é

desejado. Este dado levou assim a maioria dos trabalhos revistos na literatura e a execução deste projecto a se focar na abordagem menos complexa, a supervisionada.

Capítulo 4

Resultados

Neste capítulo são apresentados e analisados os vários resultados obtidos recorrendo à abordagem supervisionada e não supervisionada. Primeiro ponto, na secção 4.1., mostram-se os resultados obtidos através do uso das *Meta* e *Syntax features* e, de seguida, os conseguidos recorrendo aos *N-Grams*, na secção 4.2 e à análise lexical, na secção 4.3.. Por fim, conclui-se com uma apreciação critica.

4.1. Recorrendo a *Meta* e *Syntax-Features*

Na validação do classificador *Naive Bayes*, considerando os valores reais sobre a existência das *features*, as médias sobre as percentagens de classificações correctas estão descritas na tabela no Anexo J. Comparando os resultados obtidos, verifica-se que os melhores, no geral, são alcançados recorrendo às *features* selecionadas pela pesquisa exaustiva apesar das diferenças entre resultados não serem grandes. A dificuldade em classificar a neutralidade está presente em todos os casos. Verifica-se apenas que utilizando estas *features* selecionadas pelo método exaustivo, existe um aumento da classificação correcta de *tweets* positivos, no caso dos grupos contendo todos os *tweets*, sem uma perda significativa nos negativos. Consegue-se uma melhor classificação no problema da polaridade, devendo ser este grupo de *features* o selecionado. Recorrendo a este grupo de *features*, e variando o número de *tweets* para treino, obtiveram-se os resultados apresentados na Gráfico 1 para classificação de novos *tweets* no problema da subjetividade.

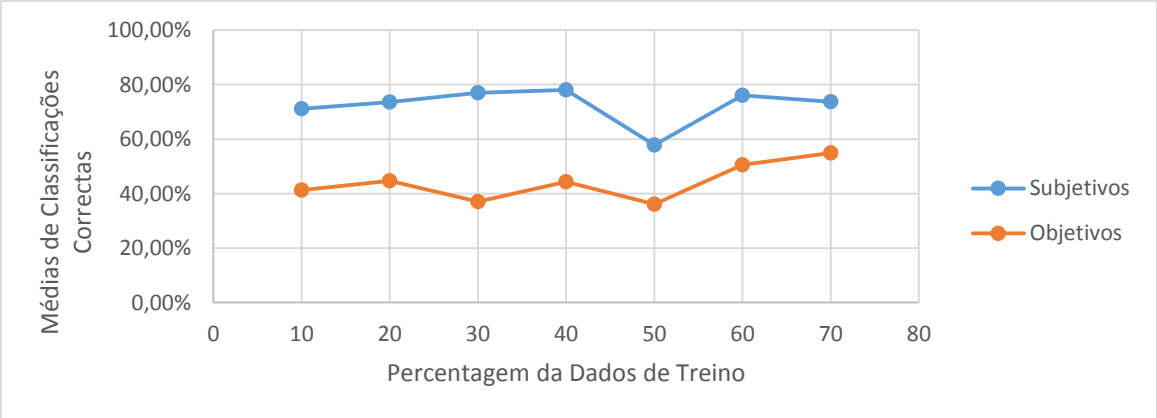


Gráfico 1: Naive Bayes, Curvas de Aprendizagem, Classificação Subjetividade

O classificador foi mantendo, de forma geral, o número de classificações correctas na sua validação com o aumento do número de dados para treino apesar de, no caso dos 50%, o número de classificações correctas tem o seu valor mais baixo para ambas as classes de *tweets*. Decidiu-se utilizar o conjunto de *features* selecionada anteriormente, mais a configuração de 70% dos *tweets* para treino no caso de classificação sobre a subjetividade. Esta escolha prende-se com a percentagem de dados onde se verifica o maior número de classificações correctas sobre a classe dos objetivos (neutros). Estre dado seria um pouco previsível, visto que este tipo de *tweets* existe em menor número nos dados disponiveis. Assim deveriam ser necessários mais dados de treino para uma boa classificação.

Para o caso da classificação sobre a polaridade, foram realizados os mesmos testes, tendo obtido os resultados apresentados na Gráfico 2.

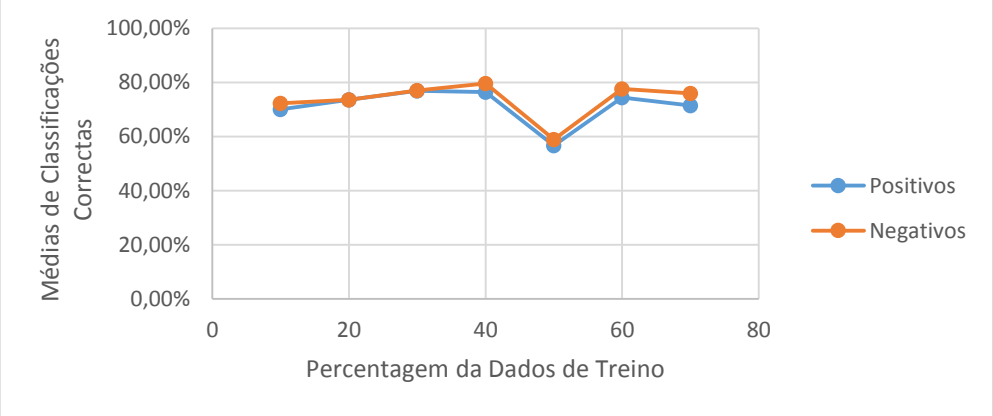


Gráfico 2: Naive Bayes, Curvas de Aprendizagem, Classificação Polaridade

Verifica-se que o comportamento deste seguiu o comportamento observado no Gráfico 1. Foi com 50% dos dados que se obtiveram os piores resultados, sendo este um dado previsível visto que se a classificação sobre a polaridade depende da classificação sobre a subjetividade, e se esta desceu neste ponto percentual, esta classificação sobre a polaridade também teria de descer. Variar para este tipo de classificação o número de *tweets* para o treino, traduziu-se numa descida das classificações correctas a partir do 40%. Querendo o máximo de classificações correctas, será esta a percentagem de dados a introduzir no treino do sistema para o problema da polaridade.

Com estes dados sobre o classificador *Naive Bayes* para os problemas de polaridade e subjetividade, foram selecionados dois conjuntos de testes a serem realizados com o classificador *Maximum Entropy* para comparação de resultados.

Num primeiro caso, realizaram-se testes recorrendo a todas as *features* sobre as características que se encontram nos *tweets*, Anexo K. Com a representação binária e

real, obtiveram-se os resultados apresentados na tabela do Anexo L. A utilização da representação real verifica-se ser a mais favorável em dois grupos, *TwitterSentiment* e Universidade Texas, obtendo-se os melhores resultados com esta representação para ambas as classificações. Ainda, o terceiro grupo (Todos os *Tweets*) necessitou desta representação para bem classificar dados quanto à subjetividade. Sabendo que todas estas dezassete *features* representam as várias características que se encontram nos *tweets,* verificou-se que os classificadores necessitaram dos valores reais sobre as contagens destas para conseguirem a melhor separação entre classes e apresentarem a melhor classificação possível.

O segundo caso prende-se com o uso das melhores *features* encontradas pelo classificador *Naive Bayes* (apresentadas na caixa de texto no Anexo M). Utilizando ambas as representações, binária e real, executam-se os respectivos testes e obtêm-se os resultados apresentados na tabela no Anexo N. A melhor representação, em geral, mostrou ser a binária. Obtiveram-se os melhores resultados recorrendo a estas *features* em três grupos, *Standford140,* TwitterSentiment e Universidade do Texas. Verificou-se que, recorrendo a informação sobre *POS-Tags*, não será importante saber o seu valor real (quantas vezes está contido no *tweet*), mas sim saber se este está contido no *tweet* ou não.

Conclui-se dos testes que ambas as representações, binária ou real, poderão ser favoráveis para se conseguir o maior número de classificações correctas. Mas que a escolha desta representação irá sempre depender dos dados que o classificador receber para o seu treino.

No primeiro grupo de *tweets*, foram obtidos melhores resultados com a representação binária, mesmo alterando as *features* a utilizar. Para o segundo, terceiro e quarto grupo verifica-se que a escolha da representação binária ou real está dependente das *features*. Para o quarto grupo, os melhores resultados foram obtidos a quando da utilização das dezassete *features* com valores binários para a classificação sobre a polaridade, e valores reais para o problema da subjetividade. Não era algo que se esperaria, mas ao pormenor os melhores resultados foram obtidos, por margem mínima, através das 17 *features*, como se poderá também verificar pelos valores sobre a sensibilidade e especificidade apresentados nos gráficos no Anexo O.

4.2. Recorrendo a *N-grams*

Considerar o uso de *N-Grams* tal como os *arrays*, contendo informação quantitativa, revelou-se ineficaz para o treino de um classificador. Também se verificou que na seleção das melhores *features* a serem utilizadas pelo classificador *Naive Bayes*, estas

nunca foram selecionadas. Isto dever-se-á a que os *N-grams* devem ser processados de forma diferente, isto é, em vez de se verificar a existência, ou quantos existem, dever-se-á verificar que *N-grams* existem.

Neste trabalho consideraram-se os grupos de *N-grams* apresentados na Tabela 1 no ponto 3.2.3.. Sabendo que as *stopwords* não contêm informação sentimental, os vários conjuntos que as contêm não foram considerados. Os grupos contendo este tipo de palavras conterão informação que deverá ser desnecessária para o tipo de classificação que se pretende, além de que mais dados para processamento se traduzem em mais tempo necessário ao treino. Também o caso do grupo *trigrams* sem *stopwords* não será considerado devido a não terem sido encontrados grupos de três palavras sem *stopwords*.

Os resultados dos testes efectuados com o classificador *Naive Bayes* são apresentados nas tabelas seguintes. A Tabela 5 apresenta os resultados médios obtidos na validação relativamente às classificações correctas para o problema da Subjetividade, e a Tabela 6 para a Polaridade.

Grupo de *Tweets*	***Sentiment140***	**Universidade Texas**	**Todos os *Tweets***
Unigrams sem Stopwords	**65,5%**	**75,7%**	67,6%
Unigrams + POS-Tags	46,0%	66,8%	51,1%
Bigrams sem Stopwords	42,7%	40,5%	40,1%
Trigrams	32,6%	48,6%	47,0%
Feature Selection Best – Unigrams sem Stopwords	53,4%	71,7%	**69,3%**

Tabela 5: Classificações correctas sobre a Subjetividade recorrendo a N-Grams

Observa-se que os *unigrams* são os mais favoráveis a uma boa classificação. Para qualquer grupo é com este tipo de *N-gram* que se obtêm os melhores resultados. Um dado interessante diz respeito ao melhor resultado do grupo de todos os *tweets*. Ao contrário dos outros grupos, apenas neste o uso das *features* mais relevantes produziu melhores resultados, tal como se verifica para todos os grupos na Tabela 6.

Grupo de *Tweets*	***Sentiment140***	***TwitterSentiment***	**Universidade Texas**	**Todos os *Tweets***
Unigrams sem Stopwords	74,18%	96,78%	95,35%	94,64%
Unigrams + POS-Tags	50,00%	49,85%	49,62%	49,89%
Bigrams sem Stopwords	50,00%	49,77%	50,00%	49,47,%
Trigrams	50,00%	48,00%	50,00%	0,00%
Feature Selection Best – Unigrams sem Stopwords	**75,24%**	**96,93**	**95,44%**	**94,75%**

Tabela 6: Classificações correctas sobre a Polaridade recorrendo a N-Grams

Neste segundo caso relativo à polaridade, verificou-se que foi através do uso de *unigrams* que se conseguiu obter os melhores resultados, e que este é um problema mais fácil para os classificadores, tal como foi visto a quando da classificação recorrendo às contagens das características dos *tweets*. As percentagens de classificações correctas são sempre maiores para este tipo de problema do que para o da subjetividade. O classificador consegue mais facilmente separar as classes positiva e negativa, do que as classes subjetiva e objetiva (neutro). Verificando a sensibilidade e especificidade para os testes do classificador *Naive Bayes,* onde foram utilizados os *unigrams*, tabelas no Anexo P e Anexo Q, verifica-se que os resultados se traduzem em valores sobre a sensibilidade (*true positive rate*) altos e, para os valores de *false positive rate* (1 – especificidade) baixos., algo desejado e positivo. Onde, mais uma vez, o caso do problema da polaridade é mais fácil de resolver visto que as diferenças entre estes valores (TPR e FPR) são maiores comparativamente com os obtidos para a subjetividade.

Utilizando os mesmos dados com o classificador *Maximum Entropy*, verificou-se o mesmo. Observando as tabelas no Anexo R, Anexo S e Anexo T, com o uso de *unigrams*, *bigrams* e *trigrams* respectivamente, verifica-se que o uso de *unigrams* é a melhor solução. Poder-se-á afirmar que o uso de *N-grams* com o classificador *Maximum Entropy* para a classificação dos *tweets* também é uma opção muito viável. Em todos os casos, obtiveram-se percentagens de classificações correctas, mesmo nos piores casos, acima dos 70%, representando bons resultados.

Entende-se através de todos estes dados o porquê de, na maioria dos trabalhos revistos, se utilizarem como *features* para classificação dos textos os *N-grams* e nunca o *array* de *features,* binário ou real.

Utilizando o grupo de *tweets* da Universidade do Texas como uma amostra dos dados disponiveis, registaram-se as curvas de aprendizagem do classificador *Naive Bayes*, recorrendo aos *unigrams* para a classificação (Gráficos 3 e 5), e sobre o classificador *Maximum Entropy* (Gráficos 4 e 6).

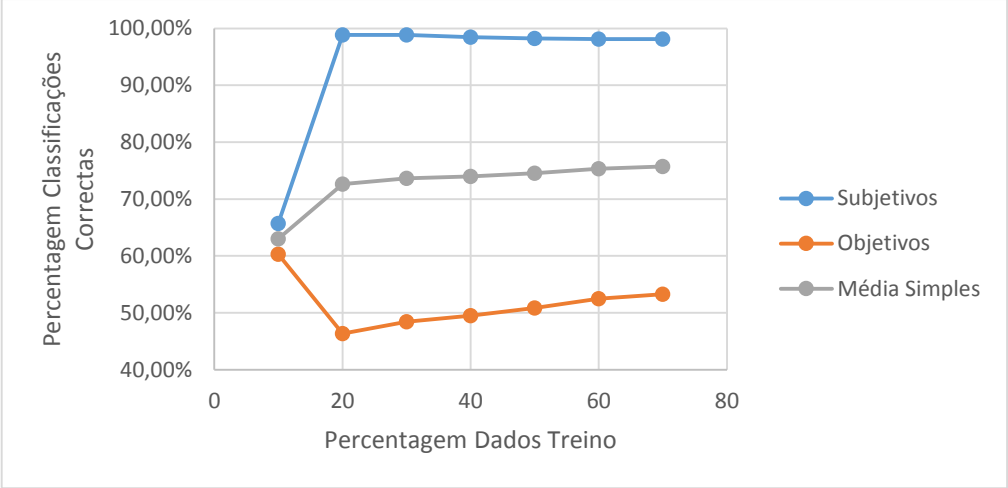


Gráfico 3: Curva Aprendizagem, Classificador Naive Bayes, Subjetividade

Pelo Gráfico 3 quanto maior é a percentagem de dados objetivos para treino, maior é o número de classificações correctas destes. Ou seja, para a aprendizagem sobre a neutralidade dos *tweets* será necessário conseguir o maior número de dados pré-classificados para se conseguir obter uma precisão cada vez mais alta. Em média as classificações correctas encontram-se entre os 60 e 80%, o que são bons resultados, mas esta média é bastante influenciada pelos valores mais altos da classificação dos *tweets* subjetivos.

Já para o classificador *Maximum Entropy*, verifica-se que o pico da classificação para os *tweets* objetivos (neutros) não se encontra nos 70%, tal como se apresenta no Gráfico 4.

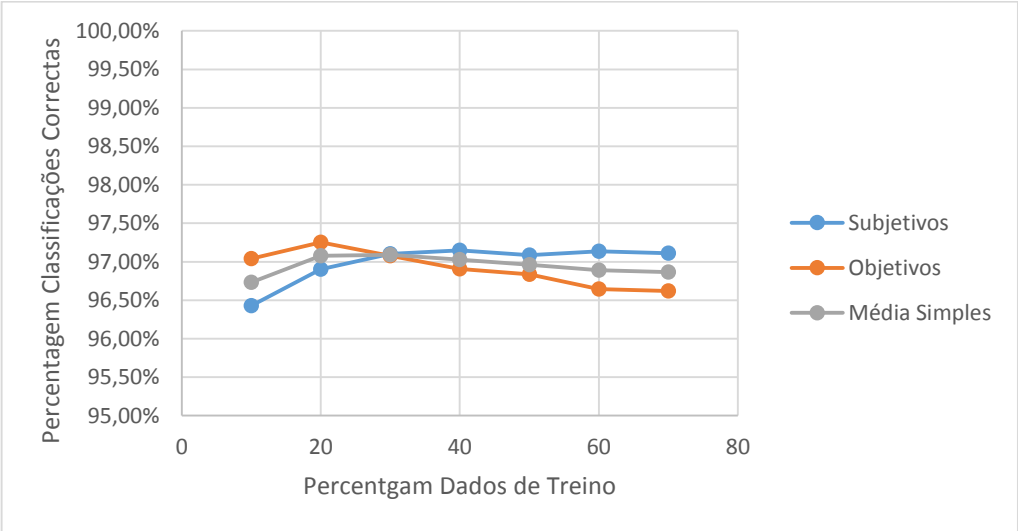


Gráfico 4: Curva Aprendizagem, Classificador Maximum Entropy, Subjetividade

Este classificador retorna o ponto mais alto para a classificação sobre os objetivos nos 20%, obtendo resultados piores a cada aumento de dados para esta classe de *tweets.*

Para o caso das classificações de *tweets* subjetivos verifica-se que a classificação melhorou até aos 40%, tendo a partir deste valor, mantido a percentagem sobre as suas classificações correctas. Para se visualizar a separação das classes, positivo e negativo para estes *tweets* subjetivos, com o classificador *Naive Bayes*, apresenta-se o Gráfico 5.

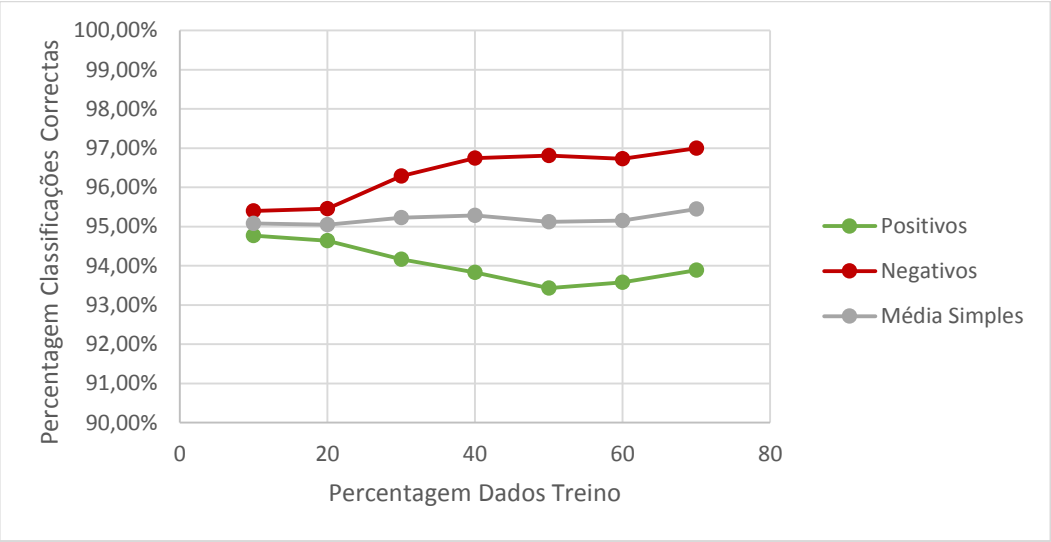


Gráfico 5: Curvas Aprendizagem, Classificador Naive Bayes, Polaridade

Neste Gráfico 5, o número de *tweets* positivos classificados correctamente desce enquanto as classificações negativas sobem. Mas, com esta situação a encontrar-se com percentagens acima dos 90%, estes representam mesmo assim resultados bons. Comprova-se também que um classificador separa estas duas classes de forma mais fácil, visto não necessitar de uma grande percentagem de dados. Logo com o mínimo, 10%, dos dados retornou boas classificações tal como também se verifica para o classificador *Maximum Entropy*, Gráfico 6.

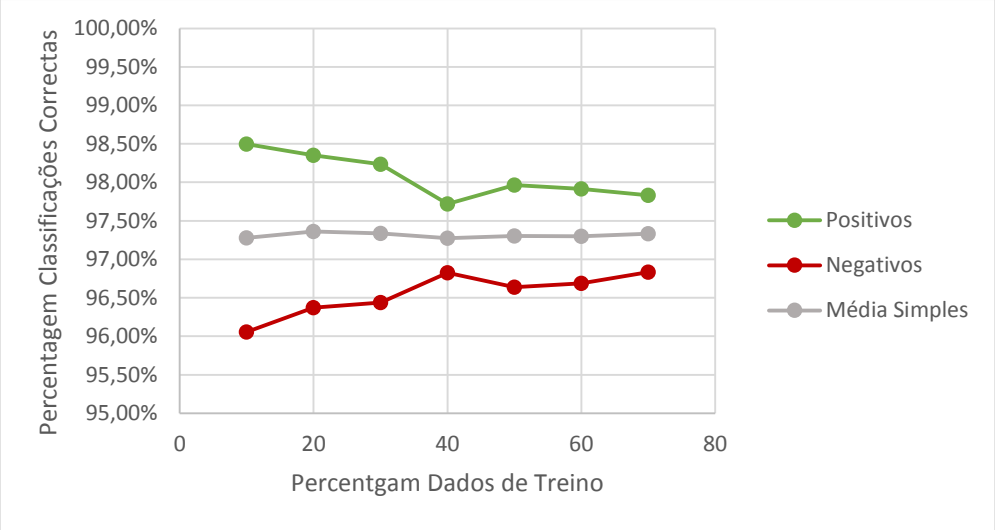


Gráfico 6: Curva Aprendizagem, Classificador Maximum Entropy, Polaridade

Com todos estes resultados pode ser concluído que, para os dados que estão disponiveis e para o classificador *Naive Bayes*, no caso da classificação da subjetividade será necessário utilizar o maior número de dados para o treino devido à classificação sobre a neutralidade. Isto é contrário do classificador *Maximum Entropy,* que recorrendo às dependências entre as *features* consegue melhores resultados no problema da neutralidade, conseguindo a partir de menos dados melhores resultados sobre a classificação da subjetividade.

A importância sobre os *N-grams* poderá ser dada principalmente devido aos resultados que se obtêm sobre o problema da subjetividade, enquanto para o caso da polaridade com poucos dados, e recorrendo a ambas as representações (vectores e *n-grams*), se conseguem boas classificações em ambos os classificadores.

Verificando os resultados da sensibilidade e especificidade obtidos pelo classificador *Naive Bayes*, apresenta-se o seguinte Gráfico 7. Com cor laranja representa-se o espaço de classificações obtido para a polaridade e com cor cinza sobre a subjetividade.

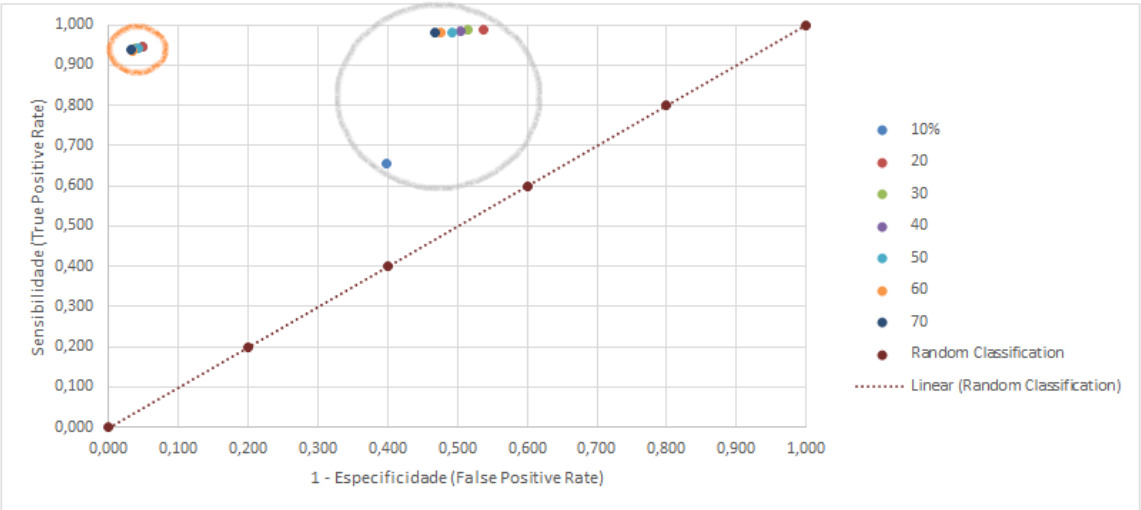
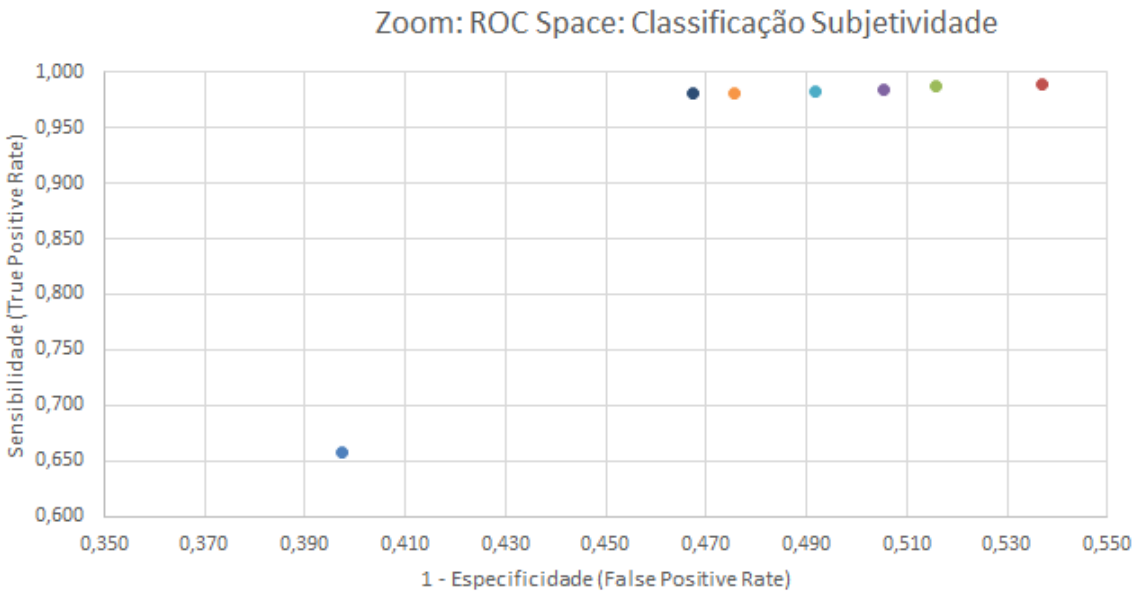


Gráfico 7: Espaço ROC, Subjetividade e Polaridade, Classificador Naive Bayes

Ambos os casos se encontram no lado positivo do gráfico, o espaço que representa resultados positivos para as classificações. A classificação da polaridade obtém melhores resultados que da subjetividade, onde se observa a percentagem sobre o rácio dos falsos positivos mais alto. Algo previsível e que melhora um pouco devido ao aumento do número de dados para treino. Mais uma vez, este valor representa a já referida dificuldade em classificar a classe objetiva e assim a dificuldade na classificação da subjetividade. Alterando os valores dos eixos do Gráfico 7, obtém-se uma aproximação aos valores resultando nos gráficos seguintes na Tabela 7.



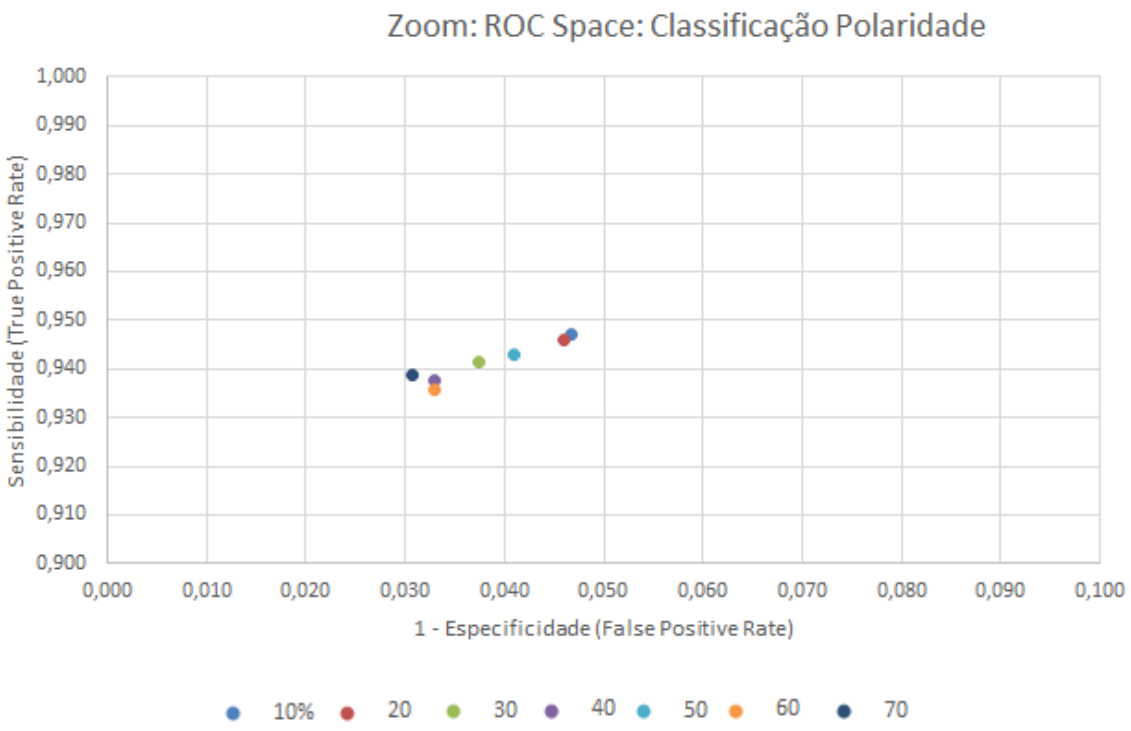


Tabela 7: Espaços ROC, Classificador Naive Bayes

Por estes gráficos da Tabela 7 descrevem-se os resultados apresentados nas tabelas no Anexo P e Anexo Q. Para a subjetividade, as diferenças entre a sensibilidade e especificidade são menores, representando piores resultados, que para o caso da classificação sobre a polaridade. O rácio dos falsos positivos para o caso da polaridade obteve um valor máximo de 0,05 enquanto para a subjetividade o valor mínimo é um pouco maior que 0,5. Os resultados obtidos podem ser considerados positivos mas existirão, como até aqui, sempre problemas relacionados com a classificação da subjetividade que deverão ser melhorados recorrendo a mais dados.

No caso do classificador *Maximum Entropy* estes também são apresentados na Tabela 8, e sendo os seus valores sobre o rácio dos falsos positivos menores, este será considerado como o melhor classificador.

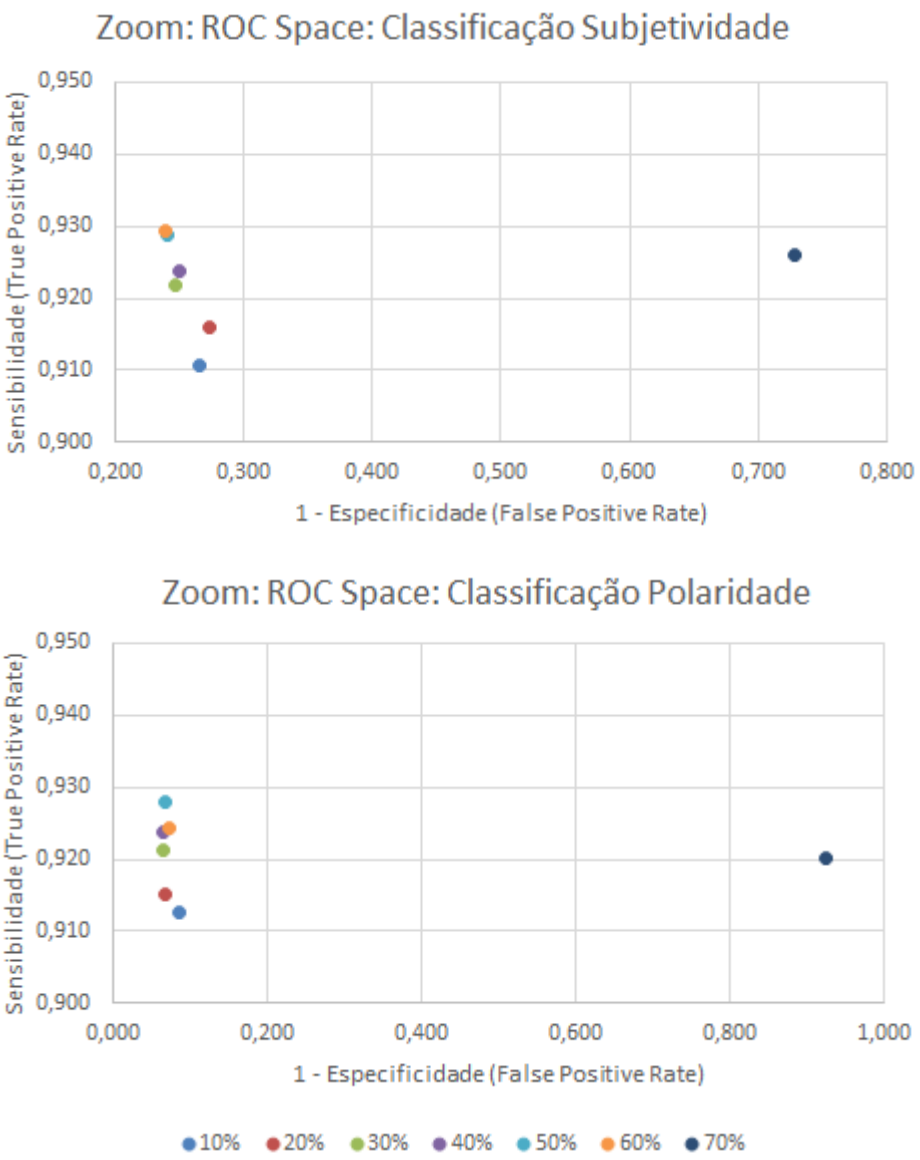


Tabela 8: Espaços ROC, Classificador Maximum Entropy

Comparando a Tabela 7 com os resultados da Tabela 8, verifica-se que no pior caso, a classificação da subjetividade, os melhores resultados são obtidos recorrendo ao classificador *Maximum Entropy.* Os valores sobre o rácio dos falsos positivos na classificação da subjetividade neste classificador no pior caso está perto do valor 0,3, enquanto na Tabela 7, para o classificador *Naive Bayes,* o seu pior valor está perto do valor 0,5. Para o caso sobre a classificação da polaridade os resultados são semelhantes. Ambos os classificadores obtiveram valores altos sobre o rácio sobre os verdadeiros positivos e um rácio dos falsos positivos mais baixo, representado um maior número de classificações correctas foi o classificador *Maximum Entropy*.

4.3. Recorrendo à Análise lexical

Tal como na obtenção de resultados recorrendo aos classificadores, alguns testes foram executados recorrendo à análise lexical dos vários grupos de *tweets.* Esta análise tem como base contagens de palavras positivas existentes no *tweet*, palavras negativas, *emoticons* positivos, e todos os demais definidos. O cálculo do sentimento positivo, negativo ou neutro baseado apenas na análise lexical, foi efectuado segundo o Pseudocódigo 2.

```
Se (#Positivo E #Negativo > #Neutro)
        Se (#Positivo > #Negativo)
                Tweet Positivo
        Senão (#Positivo < #Negativo)
                Tweet Negativo
        Senão (#Positivo == #Negativo)
                Tweet Neutro
Senão
        Tweet Neutro
```

Pseudocódigo 2: Método de Análise Lexical

Com este pseudocódigo definem-se algumas premissas que poderão ser diferentes de autor para autor em trabalhos existentes na literatura. Tendo calculado os dados referente às contagens de positivos, negativos e neutros, considerou-se para esta análise o mesmo que para os classificadores.

Em primeiro lugar deve ser realizada a verificação sobre a subjetividade. No caso do *tweet* ser considerado neutro o processo pára, mas se o *tweet* for considerado subjetivo, verifica-se a sua polaridade. Para isto os sinais de maior, menor e igual serão considerados em separado. Considerando neste trabalho os *tweets* neutros, não teria sentido definir que por exemplo, se a contagem de positivos for maior ou igual que a contagem de negativos, então o tweet é classificado como positivo. Este caso poderá fazer sentido apenas quando existem duas classes, positivo e negativo. Recorrendo a este processo os vários grupos de *tweets* foram classificados e por fim os resultados finais obtidos, são apresentados no Gráfico 8.

Gráfico 8: Espaço ROC sobre Análise Lexical

Como se verifica, os resultados não são maus. Os vários grupos de *tweets* obtiveram resultados positivos e, como também se verificou para os classificadores, piores para o problema da subjetividade (contorno a cinza) e melhores para a polaridade (contorno a laranja). Os melhores resultados para o problema da subjetividade foram os para o grupo *Stanford140.* Devido à quantidade de *tweets* neutros existente neste grupo, este dado seria previsível. Contendo o menor número de *tweets*, em comparação com os outros grupos, daria origem a um número menor de classificações erradas. No caso da classificação da polaridade, esta foi realizada com menos erros pelo grupo *TwitterSentiment*, o grupo que apenas contêm *tweets* positivos e negativos, conseguindo-se assim o maior número de classificações correctas no problema da polaridade. No grupo da Universidade de Texas, onde existe um número semelhante para as três classes de *tweets* verifica-se que os resultados foram, não os melhores, mas positivos quer na separação de classes no problema da subjetividade como na polaridade. Retira-se destes resultados que é possível, recorrendo a uma análise lexical, realizar as classificações desejadas. Tal como concluído na análise dos classificadores, a separação das classes no problema da polaridade obterá melhores resultados que na subjetividade. Além disto, por exemplo, no problema da polaridade, nesta análise os resultados do rácio dos verdadeiros positivos ficaram-se pelo valor de 0,7. Este valor está duas décimas abaixo dos conseguidos pelos classificadores no mesmo problema mostrando que apesar da classificação ser possível não será tão boa como a feita pelos classificadores.

4.4. Apreciação Crítica

Na literatura, por exemplo nos trabalhos [6], [8], [14], [17], onde se implementou um classificador *Naive Bayes*, apresentam-se resultados em média de 70% de boas classificações considerando apenas classificação sobre o problema da polaridade. O caso dos *tweets* neutros foi deixado de lado, tal como as *features* escolhidas foram baseadas apenas em *N-Grams*. Cada trabalho, cada implementação terá as suas próprias características, sendo que os valores obtidos neste trabalho são considerados como positivos e dentro da média dos trabalhos revistos, apesar do número reduzido de dados em comparação com os trabalhos da literatura. No caso da classificação recorrendo às características dos *tweets* verificou-se que a classificação sobre o problema da subjetividade de um *tweet* é difícil. Com o sistema implementado, a melhor percentagem foi nos *tweets* do grupo da Universidade de Texas com quase 60%. Esta dificuldade mostra que estes classificadores tendem a considerar muitos dos *tweets* como textos que contêm sempre algum tipo de informação positiva ou negativa. Este caso também se verificou quando se recorreu aos *N-grams,* onde os resultados melhoraram mas apenas na classificação sobre a polaridade. Para a classificação da subjetividade o classificador que recorre às características dos *tweets* mostrou-se melhor. Apesar destes resultados, deverá ser possível melhorar se existir um maior número de dados neutros para o treino dos classificadores e assim, tal como acontece para o caso da polaridade, obter melhores resultados recorrendo aos *n-grams* do que às características dos *tweets*.

Outros trabalhos, como os [6], [34] e [35], onde se implementa um classificador *Maximum Entropy*, reportam 80%, 81% e 64% de classificações correctas respectivamente. O sistema implementado neste trabalho apresentou diferenças. O classificador obteve, para os dados disponiveis, em média cerca de 85% a 95% de classificações correctas na sua validação. Este resultado é muito positivo, além de que neste trabalho se considerou a componente da classificação sobre a neutralidade, algo que na maioria dos trabalhos da literatura não é feito. Uma possível razão para os bons resultados obtidos passará pela separação dos problemas. Não foi construído um sistema que trabalha directamente sobre três classes (positivo, neutro, negativo), mas sim sobre duas de cada vez, subjetivo contra objectivo e, posteriormente, positivo contra negativo. Além disto, o próprio pré-processamento executado sobre os *tweets* poderá ser melhorado. As principais características dos *tweets* foram analisadas e tratadas, transformando os dados em bruto em bons dados de suporte à classificação. Outros factores poderão ser considerados tal como o número total de dados recolhidos, os 20.000 *tweets* pré-classificados. Este é um número pequeno comparando com os vistos

na literatura onde se utilizaram entre os 100.000 a 500.000 *tweets*. Utilizando um maior número de *tweets,* deverão ser conseguidos classificadores com uma maior generalização. Ou seja, tendo neste trabalho um menor número de dados, poderá ajudar a subir os resultados das boas classificações mas possivelmente não serão tão genéricos como os revistos. Um bom teste aos sistemas implementados neste projecto, e certamente realizado num futuro próximo, passará por se retirar directamente *tweets* através da API do Twitter (tal como nos trabalhos revistos). Utilizar o método simples de pré-classificação (uso dos *emoticons*), e verificar os resultados obtidos com os classificadores implementados. Escolhe-se este método visto que se deseja obter uma grande quantidade de *tweets* e que a classificação manual é uma tarefa complicada, não existindo recursos para ela. Como se verifica pelo projecto *Tweenator*[40] onde o autor, não tendo recursos humanos para tal, disponibilizou uma ferramenta numa página *web* mas apenas conseguiu 17% da sua lista de *tweets* classificada, até à consulta desse trabalho.

Implementados e validados ambos os classificadores, pode ser concluído que as abordagens supervisionadas são a melhor solução para o problema. Apresentando os espaços ROC na tabela seguinte, Tabela 9, sobre cada um dos classificadores finais, retira-se facilmente qual o melhor. Neste trabalho será o classificador *Maximum Entropy.*

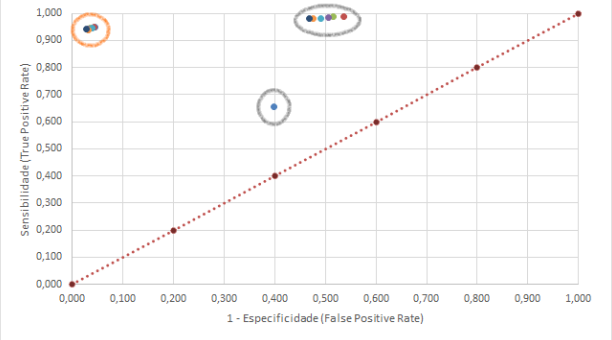
Classificador *Naive Bayes*	**Classificador *Maximum Entropy***
	

Tabela 9: Espaços ROC, Resultados Finais

Para o caso da classificação sobre o problema da polaridade (contorno a laranja) verifica-se que ambos os classificadores aprenderam e distinguiram bem as duas classes. Para o caso da classificação sobre o problema da subjetividade (contorno a cinza) verifica-se que o classificador *Maximum Entropy* obtém melhores resultados no geral. Apesar destes não serem tão bons como os obtidos na classificação da polaridade, este classificador devolve resultados que são um pouco comparáveis com

[40] Http://www.tweenator.com/

esses para alguns grupos de *tweets*, dado principal para o considerar como o melhor classificador. No caso do classificador *Naive Bayes* isto não se verificou. Este encontrou bastantes dificuldades em separar as duas classes no problema da subjetividade.

Ambos obtiveram os seus melhores resultados através do uso de *unigrams sem stopwords*. Ambos poderão ser utilizados para a classificação sentimental de *tweets* que se deseja, tal como a análise lexical mas o recomendado por este trabalho é a abordagem supervisionada e o classificador *Maximum Entropy.*

Capítulo 5

Apreciações Finais

Todo este trabalho envolveu atribuir um dado sentimento a textos com características de *Microblog*, mais concretamente os sentimentos de positivo, negativo e neutro a *tweets.* Para isto, na perspectiva de abordagem supervisionada, implementaram-se classificadores baseados em algoritmos de aprendizagem, bem como uma análise lexical referente a abordagem não supervisionada. Vários métodos foram revistos para a execução destas abordagens, concluindo que sendo a não supervisionada a mais completa, será necessário um estudo mais aprofundado da língua dos textos para se conseguirem melhores resultados. Deverá não ser considerada apenas analisado a análise lexical mas, vários pontos sobre a gramática, análise sintática, devem ser considerados, o que torna esta abordagem um problema muito mais complexo, levando à escolha da abordagem seguida ter recaído sobre a supervisionada. Recomenda-se assim o uso de classificadores para resolver este problema visto que poderão ser deixados de lado os problemas linguísticos, mas deverá existir trabalho na recolha de dados pré-classificados, *tweets* e palavras, tal como sobre os vários métodos para a criação de um bom classificador. Nestes métodos, o primeiro passo e de grande importância é o pré-processamento executado sobre os *tweets* que necessita de atenção aos detalhes, para se transformarem *tweets*, textos contendo muito ruído e mal escritos, em dados limpos para uma fácil aprendizagem pelos classificadores.

Nos trabalhos revistos foram usualmente analisados grandes quantidades de dados, resolvendo o problema da pré-classificação através do simples método que recorre a *emoticons*. Com estes conseguiram-se resultados na ordem dos 80% de classificações correctas, trabalho de *Go et al.* [6] ou no *Chaovalit e Lina* [3]. No trabalho descrito neste documento, os classificadores conseguiram resultados, em média, superiores a estes 80%, mas o número de *tweets* utilizado foi muito menor. Muitos mais dados do que aqueles disponíveis para este trabalho deverão ser tidos em conta. Talvez se obtenham valores sobre as classificações correctas em média iguais, talvez se obtenham valores mais baixos, mas tendo uma variedade de dados para o treino mais abrangente deverá ser conseguido um classificador mais abrangente, mais genérico, tornando-se assim num bom classificador seja qual for o contexto em que são escritos os textos.

Trabalho futuro

Desejando continuar com os trabalhos e melhorar mais o trabalho já desenvolvido, existem alguns aspectos que poderão ser considerados no futuro.

Um primeiro ponto prende-se com um algoritmo de aprendizagem que não foi explorado. Experimentar, analisar os resultados que se obterão recorrendo a um terceiro classificador, a *Support Vector Machine* (SVM). Esta foi abordada aquando da escrita do *background* existente para este projecto mas não existiu tempo para ser explorado.

Este último classificador encerrará o estudo das abordagens supervisionadas. Outros métodos e possibilidades prendem-se principalmente com ideias do que poderá ser conseguido recorrendo à abordagem não supervisionada, tal como detectar e/ou retirar principalmente:

a) Nomes, representando assim sujeitos, tópicos, assuntos nos textos,
b) Verbos que significarão ações, boas ou más,
c) Advérbios que usualmente são utilizados para alterar o sentido às palavras,
d) Adjetivos que qualificam outras palavras.

Através destas classes gramaticais como também pelas suas relações poderá ser possível compreender melhor qual o real sentimento que o autor do texto desejava expressar, a execução total de uma análise sintática.

Se a recolha dos textos a partir das redes sociais for realizada através de palavras-chave¸ *keywords* que poderão ser o assunto existente nos textos, deverão ser verificadas qual como o sentimento que está atribuído directamente a estas palavras. Podendo assim outro aspecto ser verificado, os subtópicos. Um estudo relacionado com subtópicos existentes nos textos e o sentimento expresso sobre estas será importante. Tal como já se referiu na escrita deste trabalho, nem sempre um sentimento expresso sobre um dado assunto tem as suas palavras escritas explicitamente nos textos. Subtópicos, conceitos, domínios entre assuntos poderão e deverão ser encontrados sendo que para isto poderá ser necessário executar um estudo focado na análise semântica, recorrendo por exemplo a ontologias.

Recolher e apresentar mais detalhes sobre os vários textos deverão ser dados com valor tanto ou mais importantes, linguisticamente, como os mais simples retirando dos textos, por exemplo, as citações. São tudo ideias a serem verificadas e analisadas, sendo que o propósito final será sempre colocar à disposição um vasto leque de possibilidades de informações que se poderão retirar de dados recolhidos através de redes sociais.

Referências

[1] Bo Pang e Lillian Lee, "Opinion mining and sentiment analysis," in *Fundations and Trends in Information Retrieval, Volume 2 Issue 1-2*, Hannover, Now Publishers Inc., 2008, pp. Pages 1-135 .

[2] Bo Pang e Lillian Lee, "Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales.," in *Proceedings of the Association for Computational Linguistics*, 2005.

[3] Pimwadee Chaovalit e Lina Zhou, "Movie Review Mining: a Comparison bteween Supervised and Unsupervised Classification Approaches," in *Proceedings of the 38th Hawaii International Conference on Systems Sciences*, 2005.

[4] Adam Bermingham e Alan Smeaton, "Classifying Sentiment in Microblogs: Is Brevity an Advantage?," in *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*, New York, 2010.

[5] Luciano Barbosa e Junlan Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010.

[6] Alec Go, Richa Bhayani e Lei Huang, "Twitter Sentiment Classification using Distant Supervision," CS224N Project Report, Stanford, 2009.

[7] Alexander Pak e Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Université de Paris-Sud, Laboratoire LIMSI-CNRS*, Orsay Cedex, France, 2010.

[8] Apporv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow e Rebecca Passonneau, "Sentiment Analysis of Twitter Data," in *LSM'11 Proceedings of the Workshop on Languages in Social Media*, 2011.

[9] Tetsuya Nasukawa e Jeonghee Yi, "Sentiment analysis: capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture*, 2003.

[10] Long Jiang, Mo Yo, Ming Zhau, Xiaohua Liu, Tiejun Zhao, "Target-dependent Twitter Sentiment Classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[11] Mary S. Neff, Roy J. Byrd e Branimir K. Boguraev, "The Talent System: TEXTRACT Arquitecture and Data Model," in *Proceedings of the HLT-NAACL 2003 Workshop*

on Software Engineering and Architecture of Language Technology systems (SEALTS), 2003.

[12] M. P. Marcus, B. Santorini e M. A. Marcinkiewicz, "Building a large annotated corpus of english: the penn treebank," in *Computational Linguistics*, 1993.

[13] Wiebe, J. e E. Rillof, "Creating subjective and objective sentence classifiers from unannotated texts.," in *Computational Linguistics and Intelligent Text Processing*, 2005.

[14] Pimwadee Chaovalit e Lina Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches," in *Proceedings of the 38th Hawaii International Conference on Systems Sciences*, 2005.

[15] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[16] Michael Speriosu, Nikita Sudan, Sid Upadhyay e Jason Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing,*, 2011.

[17] Adam Bermingham e Alan Smeaton, "Classifying Sentiment in Microblogs: Is Brevity an Advantage?," in *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*, NY, USA, 2010.

[18] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics,* vol. 22, no. 1, pp. 39-71, 1996.

[19] Kamal Nigam, John Lafferty, Andrew McCallum, "Using Maximum Entropy for Text Classification," *IJCAI-99 workshop on machine learning for information filtering,* vol. 1, pp. 61-67, 1999.

[20] Guiasu, Silviu e Abe Shenitzer, "The principle of maximum entropy," in *The mathematical intelligencer*, 1985, pp. 42-48.

[21] J.N. Darroch, D. Ratcliff, "Generalized Iterative Scaling for Log-Linear Models," in *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, 1972, pp. 1470-1480.

[22] Stephen Della Pietra, Vincent Della Pietra, John Lafferty, "Inducing Features of Random Fields," in *Patterns Analysis and Machine Intelligence, IEEE Transactions on (Volume: 19, Issue: 4)*, 1997, pp. 380-393.

[23] Nigam, Kamal, John Lafferty e Andrew McCallum, "Using Maximum Entropy for text

classification," in *IJCAI-99 workshop on machine learning for information filtering*, 1999.

[24] Claude Sammut e Geoffrey I. Webb, Encyclopedia of Machine Learning, New York: Springer Science+Business Media, 2011.

[25] Tom Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, no. 8, pp. 861-874, 2006.

[26] Theresa Wilson, Janyce Wiebe e Paul Hoffman, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Proceedings of HLT/EMNLP*, Vancouver, Canada, 2005.

[27] David A. Shamma, Lyndon Kennedy e Elizabeth F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *WSM'09 Proceedings of the first SIGMM workshop on Social media*, NY, 2009.

[28] Nicholas A. Diakopoulos e David A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, NY, 2010.

[29] James Spencer and Gulden Uchyigit, "Sentimentor: Sentiment Analysis on Twitter Data," in *In The 1st International Workshop on Sentiment Discovery from Affective Data*, 2012.

[30] M. F. Porter, "An algorithm for suffix stripping," in *Program, 14(3) pp 130−137*, 1980.

[31] JURŠIČ Matjaž, MOZETIČ Igor, ERJAVEC Tomaž, LAVRAČ Nada, "LemmaGen : multilingual lemmatisation with induced Ripple-Down rules," in *J. univers. comput. sci. vol. 16 no. 9*, 2010.

[32] Joshua Goodman, "Sequencial Conditional Generalized Iterative Scaling," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2002.

[33] Adwait Ratnaparkhi, "A Simple Introduction to Maximum Entropy Models for Natural Language Processing," Institute for Research in Cognitive Science of University of Pennsylvania, Philadelphia, 1997.

[34] B. Pang, L. Lee e S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[35] Parikh, Ravi e Matin Movassate, "Sentiment analysis of user-generated twitter updates using various classification techniques," in *CS224N Final Report*, 2009, pp. 1-18.

[36] Tom Fawcett, "ROC Graphs: Notes and Practical Considerations for Researches," in *HP Laboratories*, Palo Alto, California, 2004.

ANEXOS

Anexo A Distribuição das palavras dos *tweets* pelas classes Positiva/Negativa, Subjetiva/Objetiva

Tweets (no total são 19.779)	**Positivos (43,53%)**		**Neutros (10,52%)**		**Negativos (45,95%)**	
Total de palavras detectadas	97.936		24.188		110.252	
Total Palavras Positivas	7.780	7,94%	1.225	5,06%	5.447	4,94%
Total Palavras Negativas	4.383	4,48%	1.479	6,11%	11.081	10,05%
Total Palavras Subjetivas	7.091	7,24%	1.105	4,57%	8.031	7,28%
Total Palavras Objetivas	5.769	5,89%	1.804	7,46%	8.493	7,70%

Anexo B Negações nos *Tweets*

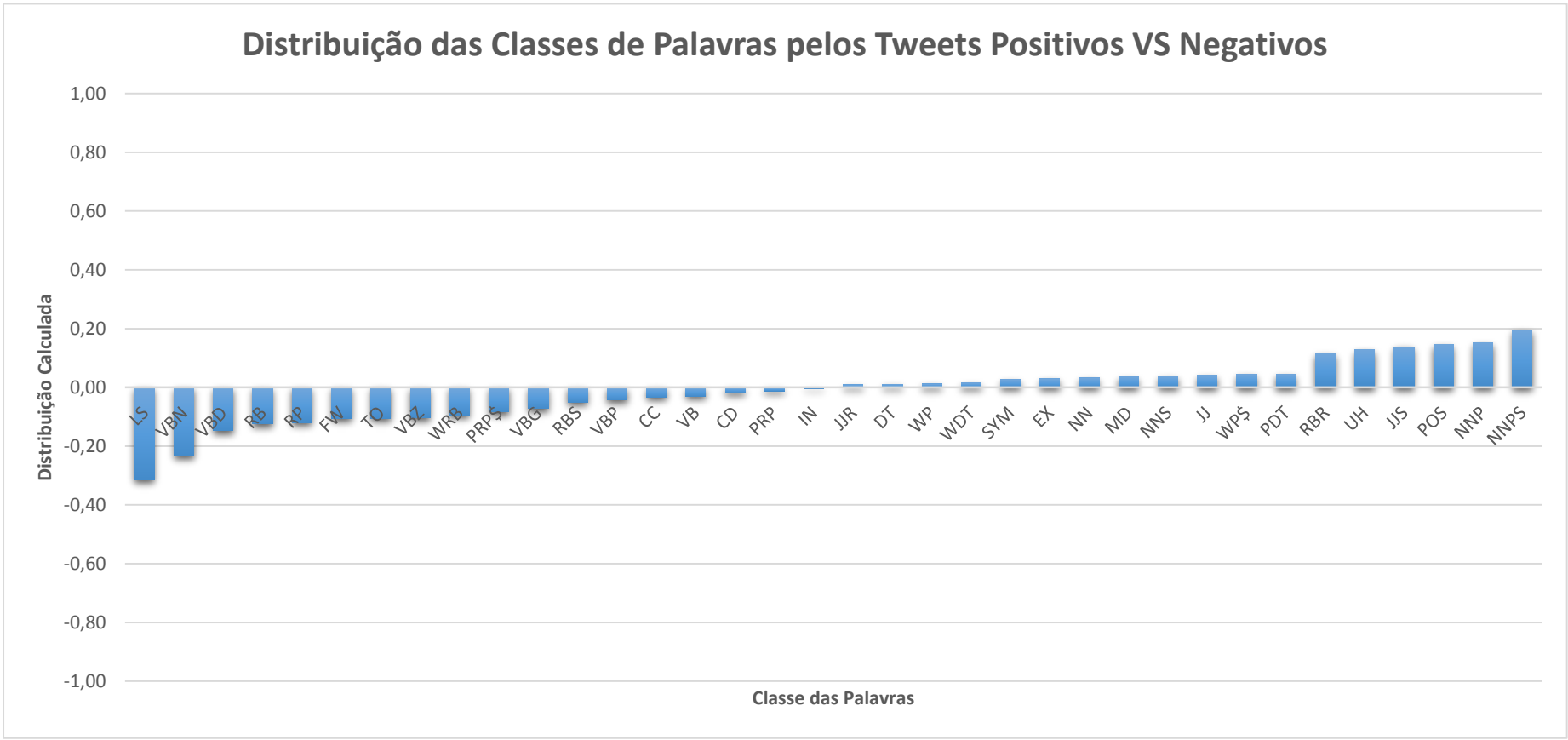
Tweets (no total são 19.779)	**Positivos (43,53%)**		**Neutros (10,52%)**		**Negativos (45,95%)**	
Negações	1265	14,69%	331	15,91%	2883	31,72%
"*Not*"	375	29,64%	107	32,33%	812	28,17%
"…n't"	**523**	41,34%	**135**	40,79%	**1305**	45,27%
"*No*"	238	18,81%	58	17,52%	532	18,45%
"*Never*"	65	5,14%	19	5,74%	128	4,44%
"*Neither*"	1	0,08%	0	0,00%	9	0,31%
"*Nobody*"	3	0,24%	2	0,60%	4	0,14%
"*None*"	6	0,47%	2	0,60%	7	0,24%
"*Nor*"	2	0,16%	0	0,00%	8	0,28%
"*Nothing*"	47	3,72%	7	2,11%	67	2,32%
"*Nowhere*"	1	0,08%	1	0,30%	5	0,17%
"*Seldom*"	0	0,00%	0	0,00%	0	0,00%
"*Hardly*"	4	0,32%	0	0,00%	6	0,21%

Anexo C Classes de Palavras e respectivos Diminutivos

Classe de Palavras	**Diminutivo**
Adverb	RB
Preposition	IN
Determiner	DT
Noun, singular or mass	NN
Possessive pronoun	PRP$
Possessive ending	POS
Verb, past tense	VBD
Palavra *"to"*	TO
Verb, base form	VB
Adjective	JJ
Proper Noun, singular	NNP
Personal pronoun	PRP
Verb, non-3rd ps. sing. Present	VBP
Verb, past participle	VBN
Noun, plural	NNS
Verb, gerund/present participle	VBG
Verb, 3rd ps. sing. Present	VBZ
Coordinating conjunction	CC
Foreign word	FW
Cardinal number	CD
"wh"-Adverb	WRB
"wh"-Pronoun	WP
Modal	MD
Symbol (mathematical ou scientific)	SYM
Interjection	UH
Adjective, comparative	JJR
Predeterminer	PDT
Particle	RP
Adverb, comparative	RBR
List item marker	LS
Adverb, superlative	RBS
Adjective, superlative	JJS
Existential *there*	EX
*"wh"-D*eterminer	WDT
Proper Noun, plural	NNPS
Possessive "*wh"*-pronoun	WP$

Anexo D Distribuição das Classes de Palavras pelos *Tweets* Subjetivos e Objetivos

Anexo E Distribuição das Classes de Palavras pelos *Tweets* Positivos e Negativos



Anexo F *Stemming* sobre as várias classes de *tweets*

Polaridade (Total de Palavras)	**Positivas (3172)**		**Neutras (570)**		**Negativas (5898)**	
Total de palavras para cálculo do Radical	2758	86,95%	430	75,44%	5139	87,13%
Total de palavras Repetidas	414	13,05%	140	24,56%	759	12,87%

Total de radicais/palavras encontradas	828	30,02%	177	41,16%	1378	26,81%
Total de radicais encontrados, palavras não	907	32,89%	127	29,53%	1466	28,53%
Total de Radicais/palavras não encontrados	1023	37,09%	126	29,30%	2295	44,66%

Anexo G *Stemming* e palavras similares

Polaridade (Total de Palavras)	**Positivas (3172)**		**Neutras (570)**		**Negativas (5898)**	
Total de palavras para cálculo do Radical	2758	86,95%	430	75,44%	5139	87,13%
Radicais encontrados, palavras não	907	32,89%	127	29,53%	1466	28,53%

Total Radicais encontrados, palavras não	343	12,44%	57	13,26%	589	11,46%
Total Radicais encontrados e palavras Similares	564	20,45%	70	16,28%	877	17,07%

Anexo H Palavras detectadas recorrendo a *Stemming* e *Lemmatization*

Tweets (total são 19.779)	**Positivos (43,53%)**		**Neutros (10,52%)**		**Negativos (45,95%)**	
Total de Palavras Positivas detectadas	7.780	7,94%	1.225	5,06%	5.447	4,94%
+ Stemming	7.780	7,94%	1.225	5,06%	5.447	4,94%
+ Lemmatization	9.841	10,05%	1.695	7,01%	6.841	6,20%
Total de Palavras Neutras detectadas	1.937	1,98%	535	2,21%	2.373	2,15%
+ Stemming	1.937	1,98%	535	2,21%	2.373	2,15%
+ Lemmatization	2.175	2,22%	626	2,59%	2.654	2,41%
Total de Palavras Negativas detectadas	4.383	4,48%	1.479	6,11%	11.081	10,05%
+ Stemming	4.383	4,48%	1.479	6,11%	11.081	10,05%
+ Lemmatization	4.695	4,79%	1.649	6,82%	11.799	10,70%
Total de Palavras Subjetivas detectadas	7.091	7,24%	1.105	4,57%	8.031	7,28%
+ Stemming	7.091	7,24%	1.105	4,57%	8.031	7,28%
+ Lemmatization	8.482	8,66%	1.356	5,61%	9.122	8,27%
Total de Palavras Objetivas detectadas	5.769	5,89%	1.804	7,46%	8.493	7,70%
+ Stemming	5.769	5,89%	1.804	7,46%	8.493	7,70%
+ Lemmatization	6.751	6,89%	2.234	9,24%	9.519	8,63%

Anexo I *Features* que poderão ser utilizadas para as classificações

***Features* para a Classificação da Subjetividade**	***Features* para a Classificação da Polaridade**
-Grupo 1: Palavras Pré-classificadas -Palavras Subjetivas, -Palavras Objetivas, -Grupo 2: Caracteristicas dos *Tweets* -Hashtags, -Retweets, -Uppercase, -URLs, -Grupo 3: Primeiras duas *POS-Tags* mais distintivas -POS, -Possessive ending -WPS, -Possessive “wh”-pronoun -Grupo 4: Segundas *POS-Tags* -UH, -Interjection -RBS, -Adverb, superlative -Grupo 5: Terceiras *POS-Tags* -FW, -Foreign word -NNPS, -Proper Noun, plural -Grupo 6: Quartas *POS-Tags* -RB, -Adverb -CD, -Cardinal number -Grupo 7: Quintas *POS-Tags* -LS, -List item marker -JJS, -Adjective, superlative -Grupo 8: Sextas *POS-Tags* -JJ, -Adjective -WP, -“wh”-Pronoun -Grupo 9: Sétimas *POS-Tags* -PRP;-Personal pronoun -NNP;-Proper Noun, singular -Grupo 10: Oitavas *POS-Tags* -VBP;-Verb, non-3rd ps. sing. Present -IN;-Preposition -Grupo 11: *Unigrams* -*Unigrams*, -*Unigrams* - *StopWords*, -Grupo 12: *Unigrams* + *POS-Tags* -*Unigrams* + *POS-Tagging*,	-Grupo 1: Palavras Pré-classificadas -Palavras Positivas, -Palavras Negativas, -Grupo 2: Emoticons -Emoticons Positivos, -Emoticons Negativos, -Grupo 3: Caracteristicas dos *Tweets* -Repetições, -Pontuação, -Grupo 4: Primeiras duas POS-Tag mais distintivas -NNPS, -Proper Noun, plural -LS, -List item marker -Grupo 5: Segundas *POS-Tags* -NNP, -Proper Noun, singular -VBN, -Verb, past participle -Grupo 6: Terceiras *POS-Tags* -POS, -Possessive ending -VBD, -Verb, past tense -Grupo 7: Quartas *POS-Tags* -JJS, -Adjective, superlative -RB, -Adverb -Grupo 8: Quintas *POS-Tags* -UH, -Interjection -RP, -Particle -Grupo 9: Sextas *POS-Tags* -RBR, -Adverb, comparative -FW, -Foreign word -Grupo 10: Sétimas *POS-Tags* -PDT, -Predeterminer -TO, -Word “to” -Grupo 11: Oitavas *POS-Tags* -WPS, -Possessive “wh”-pronoun -VBZ-Verb, 3rd ps. sing. Present -Grupo 12: *Unigrams*: -*Unigrams*, -*Unigrams* - *StopWords*,

-*Unigrams + POS-Tagging - Stopwords*, -Grupo 13: *Bigrams* -*Bigrams*, -*Bigrams - StopWords*, -Grupo 14: *Trigrams* -*Trigrams*, -*Trigrams - StopWords*	-Grupo 13: *Unigrams + POS-Tags* -*Unigrams + POS-Tagging*, -*Unigrams + POS-Tagging - Stopwords*, -Grupo 14: *Bigrams*: -*Bigrams*, -*Bigrams - StopWords*, -Grupo 15: *Trigrams*: -*Trigrams*, -*Trigrams - StopWords*

Anexo J Resultados Classificador *Naive Bayes* recorrendo às características dos *tweets*

Grupo *Tweets*	***Stanford140***	***TwitterSentiment***	**Universidade Texas**	**Todos os *Tweets***
Utilizando as *features* calculadas com o método sobre a Informação Mútua:				
Positivos	50%	85,67%	56,67%	72%
Neutros	52,33%	----------	54,33%	44,67%
Negativos	75%	98%	76,67%	87,33%
Utilizando as *features* calculadas com o método *Chi Square*:				
Positivos	43%	86,67%	52,67%	70%
Neutros	56,33%	----------	57 %	46,67%
Negativos	65,67%	98%	73%	85,33%
Utilizando as *features* calculadas recorrendo procura exaustiva:				
Positivos	51,67%	91,67%	61,67%	**78%**
Neutros	54%	----------	57,33%	**46,33%**
Negativos	52%	98,67%	71,33%	**85%**

Anexo K *Features* representantes da informação contextual de *Tweets*

a) - Acrónimos
b) - *Stopwords*
c) - *Emoticons* positivos, negativos
d) - *Uppercase*
e) - Retweet
f) - Url
g) - Repetições de letras
h) - *Usernames*
i) - *Hashtags*
j) - Pontuação (“!” ou “?”)
k) – Palavras positivas, neutras e negativas
l) - Palavras subjetivas, objetivas
m) - Negações

Anexo L Resultados médios, representação binária VS real, utilizando 17 *Features*

Grupo de *Tweets*	***Stanford140***		***TwitterSentiment***	**Universidade Texas**		**Todos os *Tweets***	
Tipo Classificação	*a)*	*b)*	*b)*	a)	b)	a)	b)
Representação recorrendo a valores Binários:							
Classificação Correcta	**83,88 %**	**77,02 %**	97,83%	96,63%	97,28%	93,42%	**95,87 %**
Classificação Errada	**16,11 %**	**22,93 %**	2,16%	3,03%	2,71%	6,57%	**4,12%**
Sensibilidade	**0,869**	**0,798**	0,969	0,976	0,975	0,976	**0,938**
Especificidade	**0,254**	**0,247**	0,013	0,054	0,029	0,340	**0,019**
Representação recorrendo a valores Reais:							
Classificação Correcta	81,67%	72,17%	**97,88%**	**96,65 %**	**97,31 %**	**94,31 %**	95,78%
Classificação Errada	18,32%	27,82%	**2,11%**	**3,34%**	**2,68%**	**5,68%**	4,21%
Sensibilidade	0,861	0,733	**0,969**	**0,976**	**0,968**	**0,978**	0,935
Especificidade	0,301	0,286	**0,012**	**0,052**	**0,022**	**0,297**	0,017

a) Classificação sobre a Subjetividade

b) Classificação sobre a Polaridade

Anexo M *Features* representantes de *Tweets* selecionadas no estudo sobre o Classificador *Naive Bayes*

I. Para a classificação da Subjetividade:
   a. O grupo 2: *Hashtags*, *Retweets*, *Uppercase* e *URLs*,
   b. O grupo 7: *POS-Tag* LS e JJS,
   c. O grupo 9: *POS-Tag* PRP e NNP,
   d. O grupo 10: *POS-Tag* VBP e IN.

II. Para a classificação da Polaridade:
   a. O grupo 1: Palavras com polaridade (positivas, negativas),
   b. O grupo 2: Emoticons com polaridade (positivos, negativos),
   c. O grupo 9: *POS-Tag* RBR e FW,
   d. O grupo 10: *POS-Tag* PDT e TO,
   e. O grupo 11: *POS-Tag* WPS e VBZ

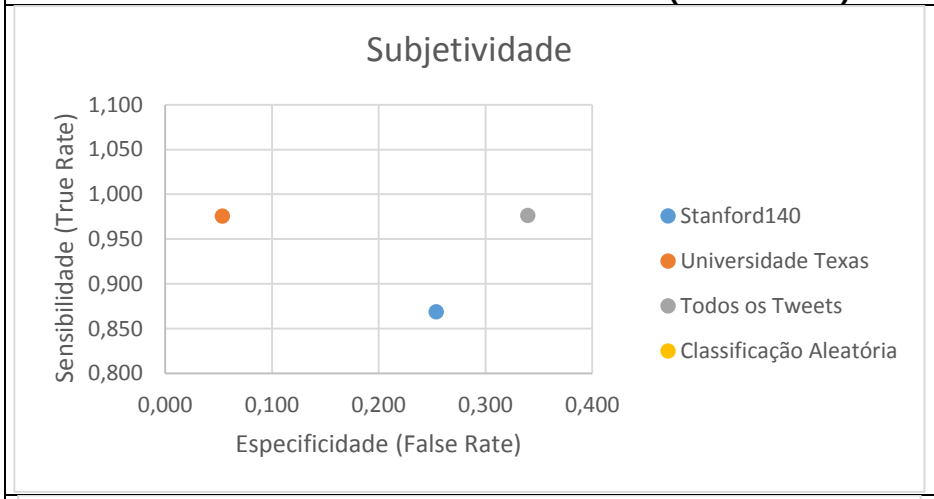
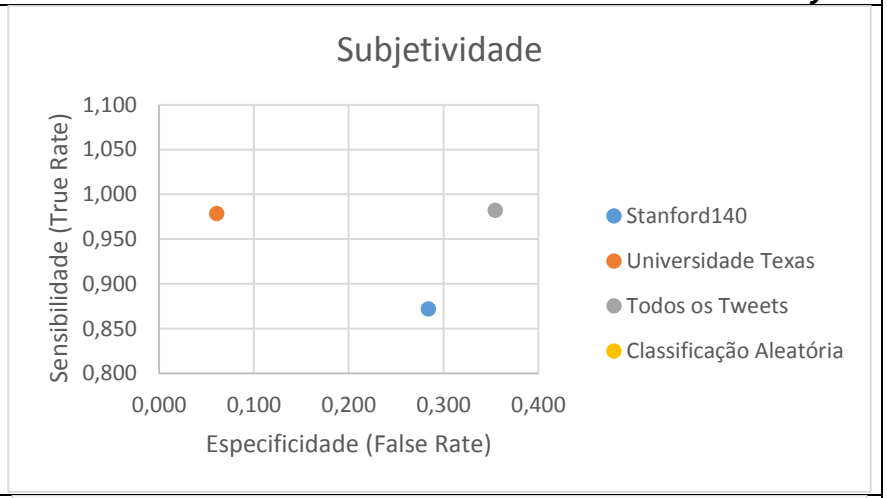
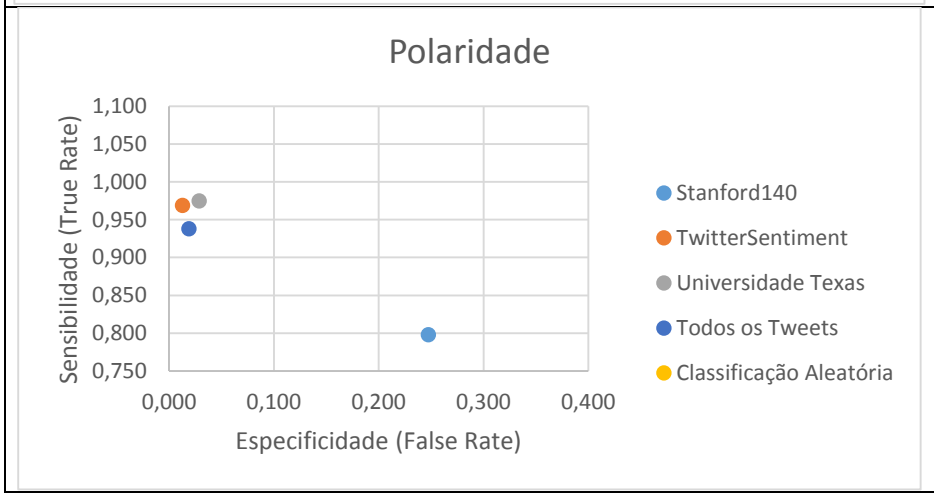
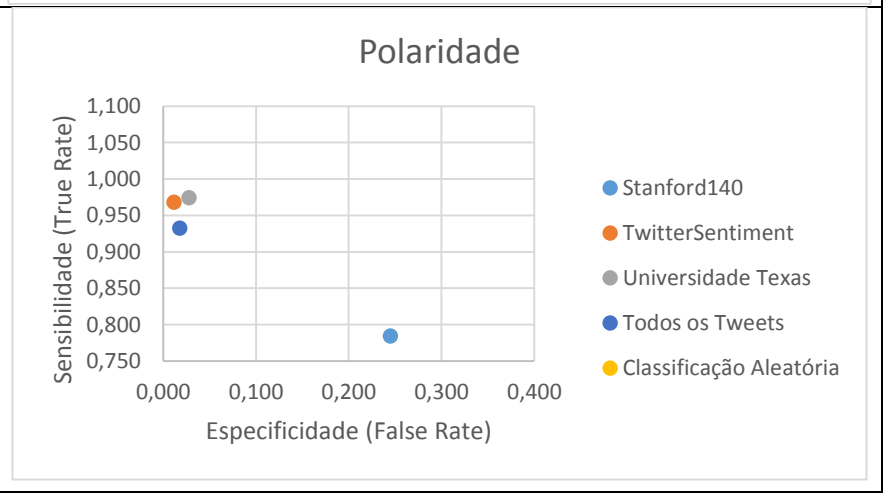
Anexo N Resultados médios, representação binária VS real, utilizando as melhores *Features* do Classificador *Naive Bayes*

Grupo de *Tweets*	***Stanford140***		***TwitterSentiment***	**Universidade Texas**		**Todos os *Tweets***	
Tipo Classificação	*a)*	*b)*	*b)*	a)	b)	a)	b)
Representação recorrendo a valores Binários:							
Classificação Correcta	**83,00 %**	**76,75 %**	**97,83%**	**96,52 %**	**97,31 %**	93,33%	95,62%
Classificação Errada	**16,99 %**	**23,24 %**	**2,16%**	**3,47%**	**2,68%**	6,66%	4,37%
Sensibilidade	**0,872**	**0,784**	**0,968**	**0,978**	**0,974**	0,982	0,932
Especificidade	**0,284**	**0,245**	**0,012**	**0,061**	**0,028**	0,354	0,018
Representação recorrendo a valores Reais:							
Classificação Correcta	80,57%	74,31%	97,82%	96,50%	97,31%	**93,61 %**	**95,73 %**
Classificação Errada	19,42%	25,68%	2,17%	3,49%	2,68%	**6,38%**	**4,26%**
Sensibilidade	0,848	0,745	0,967	0,978	0,969	**0,972**	**0,932**
Especificidade	0,321	0,259	0,011	0,061	0,022	**0,320**	**0,015**

a) Classificação sobre a Subjetividade

b) Classificação sobre a Polaridade

Anexo O Resultados Classificador *Maximum Entropy* recorrendo às características dos *tweets*

Resultados – Caracteristicas dos *Tweets* (17 *Features*)	**Resultados – Melhores Features obtidas com o *Naive Bayes***
	
	

Anexo P Especificidade e Sensibilidade, Classificador *Naive Bayes,* Subjetividade

Grupo de *Tweets*	***Sentiment140***		**Universidade Texas**		**Todos os Tweets**	
Problema Classificação	a)		a)		a)	
Variável	1)	2)	1)	2)	1)	2)
Feature Search - Unigrams sem Stopwords	**0,540**	**0,850**	**0,467**	**0,981**	0,631	0,980
Feature Selection - Unigrams sem Stopwords	0,929	0,997	0,554	0,988	**0,581**	**0,967**

a) Classificação sobre a Subjetividade

1) 1 – Especificidade (False Positive Rate)

2) Sensibilidade (True Positive Rate)

Anexo Q Especificidade e Sensibilidade, Classificador *Naive Bayes*, Polaridade

Grupo de *Tweets*	***Sentiment140***		***TwitterSentiment***		**Universidade Texas**		**Todos os *Tweets***	
Problema Classificação	a)		a)		a)		a)	
Variável	1	2	1	2	1	2	1	2
Feature Search - Unigrams sem Stopwords	0,278	0,763	0,017	0,951	0,041	**0,947**	0,042	0,934
Feature Selection - Unigrams sem Stopwords	**0,235**	**0,739**	**0,013**	**0,951**	**0,031**	**0,939**	**0,038**	**0,932**

a) Classificação sobre a Polaridade

1) 1 – Especificidade (False Positive Rate)

2) Sensibilidade (True Positive Rate)

Anexo R Resultados médios de validação do classificador *Maximum Entropy* recorrendo a *Unigrams*

Grupo de *Tweets*	***Stanford140***		***TwitterSentiment***	**Universidade Texas**		**Todos os *Tweets***	
Tipo Classificação	*a)*	*b)*	*b)*	*a)*	*b)*	*a)*	*b)*
Unigrams							
Classificação Correcta	78,36%	**77,67%**	**98,08%**	**97,01%**	**97,33%**	91,60%	95,69%
Classificação Errada	21,63%	**22,32%**	**1,91%**	**2,98%**	**2,66%**	8,39%	4,30%
Sensibilidade	0,816	**0,802**	**0,973**	**0,983**	**0,969**	0,971	0,938
Especificidade	0,349	**0,243**	**0,012**	**0,056**	**0,022**	0,421	0,023
Unigrams – Stopwords							
Classificação Correcta	**79,03%**	77,36%	98,07%	96,95%	**97,33%**	**91,65%**	**95,70%**
Classificação Errada	**20,96%**	22,63%	1,92%	3,04%	**2,66%**	**8,34%**	**4,29%**
Sensibilidade	**0,824**	0,800	0,973	0,983	**0,969**	**0,971**	**0,938**
Especificidade	**0,339**	0,246	0,012	0,056	**0,022**	**0,419**	**0,023**
Unigrams + POS-Tag							
Classificação Correcta	77,46%	75,22%	68,54%	74,07%	64,34%	88,18%	66,56%
Classificação Errada	22,51%	24,77%	31,45%	25,92%	35,65%	11,81%	33,43%
Sensibilidade	0,806	0,737	0,680	0,775	0,641	0,919	0,661
Especificidade	0,354	0,230	0,310	0,361	0,354	0,569	0,330
Unigrams + POS-Tag – Stopwords							
Classificação Correcta	77,92%	75,84%	67,77%	74,03%	63,66%	88,69%	65,78%
Classificação Errada	22,07%	24,15%	32,22%	25,96%	36,33%	11,30%	34,21%
Sensibilidade	0,811	0,751	0,669	0,771	0,633	0,916	0,652
Especificidade	0,348	0,233	0,315	0,356	0,358	0,550	0,337

a) Classificação sobre a Subjetividade

b) Classificação sobre a Polaridade

Anexo S Resultados médios de validação do classificador *Maximum Entropy* recorrendo a *Bigram*

Grupo de *Tweets*	***Stanford140***		***TwitterSentiment***	**Universidade Texas**		**Todos os *Tweets***	
Tipo Classificação	*a)*	*b)*	*b)*	*a)*	*b)*	*a)*	*b)*
Bigrams							
Classificação Correcta	**75,44%**	**74,85%**	**90,04%**	**80,07%**	**85,21%**	**89,98%**	**88,01%**
Classificação Errada	**24,55%**	**25,14%**	**9,95%**	**19,92%**	**14,78%**	**10,01%**	**11,98%**
Sensibilidade	**0,783**	**0,762**	**0,885**	**0,827**	**0,835**	**0,919**	**0,860**
Especificidade	**0,394**	**0,262**	**0,084**	**0,264**	**0,128**	**0,441**	**0,098**
Bigrams – Stopwords							
Classificação Correcta	72,78%	66,86%	80,29%	75,47%	76,21%	89,97%	78,64%
Classificação Errada	27,21%	33,13%	19,70%	24,52%	23,78%	10,02%	21,35%
Sensibilidade	0,738	0,683	0,853	0,758	0,712	0,912	0,834
Especificidade	0,426	0,312	0,230	0,264	0,154	0,433	0,246

a) Classificação sobre a Subjetividade

b) Classificação sobre a Polaridade

Anexo T Resultados médios de validação do classificador *Maximum Entropy* recorrendo a *Trigrams*

Grupo de *Tweets*	***Stanford140***		***TwitterSentiment***	**Universidade Texas**		**Todos os *Tweets***	
Tipo Classificação	*a)*	*b)*	*b)*	*a)*	*b)*	*a)*	*b)*
Trigrams							
Classificação Correcta	71,77%	66,66%	81,92%	73,30%	75,26%	89,59%	79,55%
Classificação Errada	28,22%	33,33%	18,07%	26,69%	24,73%	10,40%	20,44%
Sensibilidade	0,734	0,662	0,790	0,740	0,718	0,904	0,764
Especificidade	0,526	0,326	0,149	0,301	0,198	0,457	0,169

a) Classificação sobre a Subjetividade

b) Classificação sobre a Polaridade

Anexo U Visão Geral sobre a execução do Projecto