UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE INFORMÁTICA

Sensor de Reputação Online: técnicas de aprendizagem automática para a deteção e classificação de opiniões na Web

Jorge Miguel Ferreira Letras

Orientação: José Miguel Gomes Saias

Mestrado em Engenharia Informática

Dissertação

Évora, 2014

UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE INFORMÁTICA

Sensor de Reputação Online: técnicas de aprendizagem automática para a deteção e classificação de opiniões na Web

Jorge Miguel Ferreira Letras

Orientação: José Miguel Gomes Saias

Mestrado em Engenharia Informática

Dissertação

Évora, 2014

Sumário

As redes sociais são plataformas em larga escala onde pessoas de todo o mundo se podem conhecer, partilhar imagens e vídeos ou trocar opiniões. Saber as opiniões dos utilizadores que podem afetar a reputação de um produto ou serviço é uma das vantagens que as empresas podem retirar deste tipo de plataformas. O objetivo deste trabalho é apresentar um sistema com a capacidade de determinar, através de técnicas de aprendizagem automática, o sentimento de uma frase e respetivo impacto na afetação da reputação da entidade mencionada, classificando-o como positivo, negativo ou neutro. Este sistema foi desenvolvido na linguagem Python e utiliza recursos da ferramenta NLTK, como o reconhecimento de entidades (NE_Chunk), o classificador gramatical (pos_tag) e os algoritmos para o classificador da polaridade de sentimentos (Naive Bayes, Decision Trees e SVM).

Online Reputation Sensor: machine learning techniques for detection and classification of opinions in Web textual sources

Abstract

The social networks are large scale platforms where people around the world meet, share photos and videos and share opinions. Knowing people's opinions about a product or service is one of the advantages that companies can benefit from these type of plataforms. The purpose of this work is to present a system with the ability to predict, through machine learning techniques, the sense of a sentence and the respective reputation impact on the target entity, classifying it as negative, positive or neutral. This system was developed in Python and uses resources from NLTK framework, such as entity recognition (NE_Chunk), the grammar classifier (pos_tag) and the algorithms used in system development (Naive Bayes, Decision Trees, and SVM).

Para a minha família, namorada e amigos.

Agradecimentos

Deixo aqui uma dedicatória especial a todos aqueles que, de uma forma ou de outra, deram o seu contributo na realização deste trabalho. Sem a ajuda deles a concretização do mesmo seria bastante mais difícil.

Em primeiro lugar quero deixar expresso o agradecimento à minha família e namorada, pela motivação, apoio e confiança que depositaram em mim. O apoio deles foi fundamental não só durante este trabalho mas também em todas as etapas do meu percurso académico.

Ao meu orientador, Professor José Saias, um agradecimento especial pela sua disponibilidade, esclarecimentos, ajuda e conselhos dados ao longo de todo o trabalho. O interesse aliado aos elevados conhecimentos do Professor nesta área, foram dois dos principais contributos para que eu me sentisse sempre motivado e confiante no projeto que estava a realizar.

Ao meu colega de curso, Pedro Roque quero agradecer o companheirismo e cumplicidade. Foram muitas as horas que passámos juntos a realizar trabalhos e a estudar e, grande parte dos resultados obtidos em diferentes cadeiras, devem-se também à sua contribuição.

Agradeço também a todos os restantes colegas e professores da Universidade de Évora que, através dos seus conhecimentos e disponibilidade, contribuíram para a minha aprendizagem e formação. Os conhecimentos adquiridos ao longo de todo o curso muito me ajudaram na realização deste trabalho e também contribuirão, seguramente, para o sucesso da minha vida profissional e pessoal.

Acrónimos

URL *Uniform Resource Locator*

SVM *Support Vector Machines*

NLTK *Natural Language Toolkit*

PLN Processamento de Linguagem Natural

IA Inteligência Artificial

CLEF *Conference and Labs of the Evaluation Forum*

POS *Part-Of-Speech*

DAL *Dictionary of Affect in Language*

BoW *Bag-of-Words*

DT *Decision Trees*

NB *Naive Bayes*

TF-IDF *Term Frequency–Inverse Document Frequency*

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Enquadramento e motivação

O sucesso das redes sociais juntamente com as novas ferramentas e utilizações da *Web 2.0* causaram algumas alterações na forma como as pessoas comunicam e partilham informação entre si. Atualmente, existem diversas plataformas *online* que permitem a partilha de todo o tipo de informação entre utilizadores de todo o mundo, como os *blogs*, *micro-blogs*, redes sociais, serviços de análises a produtos e serviços, entre outros. Para além disso, com o surgimento de novas e mais desenvolvidas soluções de dispositivos móveis, muito se tem vindo a alterar no que respeita à forma e necessidade de as pessoas estarem constantemente ligadas à rede global. Num estudo feito pela empresa Nielsen[1], concluiu-se que entre 2011 e 2012 houve um aumento de 37% no tempo total anual despendido na utilização de redes sociais. Este aumento traduz-se em 121 biliões de minutos para utilizadores Americanos. A utilização de dispositivos móveis para aceder a este tipo de plataformas está a ganhar cada vez mais adeptos, havendo um aumento no acesso via *smartphones*, *tablets* e até televisões.

A maior parte das pessoas que utilizam um ou outro destes diferentes serviços fazem-no apenas de uma forma lúdica. No entanto, existem plataformas sociais em que a sua utilização é feita com um objetivo estritamente profissional. No início, um sistema de rede social era maioritariamente utilizado para a comunicação entre amigos, partilha de fotos e pouco mais. Com o aumento de funcionalidades e utilizadores, as empresas viram as imensas potencialidades destas plataformas e seguiram também a tendência, passando a estarem também presentes e adoptando estratégias de divulgação dos seus produtos e serviços. Para além de poderem chegar a uma quantidade elevada de público-alvo, as

[1]Mais informações em: http://blog.nielsen.com/nielsenwire/social/2012/

empresas têm ao seu dispor um canal de publicidade que lhe pode gerar uma grande rendibilidade de uma forma mais rápida, eficaz e de menor custo.

A informação presente neste tipo de redes pode variar entre informação pessoal através de texto, vídeos e imagens em conversas casuais entre amigos ou ter um aspecto mais formal como omitir opiniões, críticas e sugestões a um produto ou serviço prestado por uma empresa ou pessoa. A forma como os utilizadores emitem essas opiniões pode ser bastante diversificada e uma análise mais cuidada a essa informação pode indicar, muitas vezes, sentimentos diversos como o medo, a angústia, a insatisfação ou contentamento. Cada comentário introduzido neste tipo de plataformas pode ter milhares de visualizações num curto espaço de tempo e dependendo do seu conteúdo, pode afectar positivamente ou negativamente a opinião de várias pessoas quanto a um determinado assunto e afetar a reputação de um determinado serviço ou empresa.

Com a quantidade e diversidade deste tipo de conteúdo, podem-se desenvolver análises diversificadas acerca da opinião de utilizadores, sobre vários domínios, como político, económico ou social. As análises retiradas a partir dessa informação podem ter bastante utilidade, quer para delinear políticas de mudança numa empresa, esclarecer consumidores insatisfeitos ou analisar o sucesso de eventos, produtos ou até de personalidades políticas em alturas de eleições.

1.2 Objetivos e contribuições

O objetivo deste trabalho é apresentar um inovador sistema de reputação de entidades, baseado em técnicas de aprendizagem automática. A partir de um conjunto de mensagens, obtidas na rede social Twitter[2], este sistema analisa a estrutura morfológica das frases e determina o sentimento geral que as mesmas demonstram. Outra das funcionalidades é determinar se na mensagem analisada se encontra algum tipo de entidade, seja marca, produto ou nome de personalidade conhecida, e determinar se o sentimento afeta a reputação da entidade mencionada. O sistema é desenvolvido em Python[3] e utiliza ferramentas de tratamento e análise de texto do pacote NLTK[4].

1.3 Organização da dissertação

A estrutura deste trabalho está disposta da maneira a seguir indicada. No capítulo 2 vão ser apresentados exemplos de trabalhos efetuados que refletem o estado da arte em relação à mineração de dados em contexto de análise de sentimentos. A maioria dos trabalhos até esta data têm como objectivo a análise de sentimentos em opiniões de pessoas a partir de

[2] http://www.twitter.com/
[3] http://www.python.org/
[4] http://nltk.org/

grandes bases de texto presentes em redes sociais como Twitter e Facebook[5] ou páginas de análises de produtos ou serviços como TripAdvisor[6] e Amazon[7].

Em 3, são apresentadas as ferramentas utilizadas no desenvolvimento do sistema.

No capítulo 4, é feita a descrição de todo o sistema final bem como de todas as suas fases intermédias. Neste capítulo também será descrito o conjunto de dados utilizado bem como os passos de pré-processamento a que foi sujeito.

No capítulo 5, vão ser apresentados os resultados que foram obtidos pelo sistema na sua versão final com dados de teste. Será também apresentado o resultado do sistema através de um teste com um diferente conjunto de dados.

Finalmente, no capítulo 6, é apresentada a conclusão e abordadas sugestões de trabalho futuro com o objetivo de melhorar a eficácia do sistema.

[5] http://www.facebook.com/
[6] http://www.tripadvisor.com/
[7] http://www.amazon.com/

Capítulo 2

Trabalho relacionado

A área de análise de sentimentos, inserida no domínio de processamento de linguagem natural (PLN), é uma área cada vez mais estudada tem vindo, cada vez mais, a ganhar a sua importância, não só a nível académico mas também social. Um dos objetivos deste capítulo é dar a conhecer o porquê deste elevado interesse e apresentar o contexto da área da análise de sentimentos. São abordados os primeiros passos dados e o processo evolutivo até se chegar às metodologias e técnicas mais utilizadas hoje em dia. Muitas dessas metodologias, técnicas e ferramentas apresentadas neste capítulo, foram utilizadas no desenvolvimento do sistema desenvolvido no âmbito deste trabalho.

2.1 O que é a análise de sentimentos?

Este é um tema cada vez mais em voga que em muito deve ao sucesso das redes sociais e de páginas de análises de produtos. As pessoas utilizam esses espaços não só para comunicarem entre si mas também para emitirem opiniões e sentimentos acerca de vários domínios tais como produtos ou filmes. A universalidade e a acessibilidade a essa informação tornam-na numa matéria que cada vez mais é objeto de análise de forma a ser possível compreender, de uma forma automática, aquilo que as pessoas pensam. Ao longo dos últimos anos, o domínio referente ao processamento de linguagem natural sofre um grande avanço.

Antes do ano 2000 haviam poucos estudos nesta área, sobretudo devido ao fraco acesso ao conteúdo. Do ano 2000 até aos dias de hoje, foram enormes os desenvolvimentos nesta área e a granularidade da análise sobre a informação passou de documentos inteiros para se basear detalhadamente em frases ou palavras presentes num documento. No

início, começou-se a classificar documentos ou comentários num todo[31][19], por exemplo, atribuindo um sentimento geral a um documento ou a um bloco de texto, constituído por múltiplas frases ou palavras, consoante a soma das polaridades (negativas ou positivas) das palavras que o constituíam. A evolução foi seguir para a classificação por frase[12][1], uma forma mais específica, e por isso, mais difícil de analisar.

A extracção de sentimento, normalmente, tem como objectivo retornar a orientação geral de um bloco de texto baseando-se nos termos que o compõem. No entanto, em determinados contextos, seria vantajosa a obtenção de mais informação, nomeadamente a satisfação do utilizador face às diferentes características do produto. Por exemplo, a satisfação do utilizador face a um determinado telemóvel não espelha a satisfação em relação às diversas características do mesmo. Actualmente, alguns dos trabalhos desenvolvidos não se focam apenas no cálculo da polaridade global mas também tentar identificar os diferentes alvos e diferentes polaridades referidas nessas mesmas opiniões. Assim, a apreciação global demonstrada por um telemóvel pode ser positiva apesar de se ter apontado críticas á sua bateria ou vice-versa. Atendendo a este facto, vários trabalhos, como [11], propõem a abordagem à análise de sentimentos baseados em características.

2.2 Análise de sentimentos e sensor de reputação

Tecnicamente, as operações sobre a informação que foram referidas no capítulo 2.1, são as mesmas que vão ser utilizadas neste sistema. No entanto, um sensor de reputação deve ter um alvo, ou seja, uma entidade. A entidade pode ser uma marca, um produto ou uma pessoa e, na maior parte das vezes, uma frase que demonstra qualquer tipo de sentimento tem como alvo uma dessas entidades presente no texto.

A partir do referido anteriormente, um sensor de reputação deve ter duas funcionalidades principais: a análise de sentimento e a deteção da entidade. A análise de sentimento tem como objetivo a deteção da polaridade global da frase. A deteção da entidade refere-se ao alvo ao qual essa frase e respetivo sentimento dizem respeito. Para um sensor de reputação de entidades essas duas funcionalidades devem estar implementadas, caso contrário, não é retornada a informação completa.

Algumas das caraterísticas e abordagens implementadas no sistema apresentado têm como base ideias e sugestões presentes nos RepLab[1] 2012 e 2013. A sigla RepLab, diz respeito a uma tarefa incorporada no *"Conference and Labs of the Evaluation Forum"* que é um fórum realizado anualmente onde são apresentadas e discutidas novas técnicas e abordagens no estudo de análise da informação. A RepLab faz parte do fórum e diz respeito à apresentação de sistemas com o objetivo de gerir a reputação de entidades com base na análise de mensagens em redes sociais.

Esta tarefa tem um cariz de competição na medida em que vários sistemas apresentados

[1] http://www.limosine-project.eu/events/replab2013

```
arg1 RELATION_NAME arg2      domain;register;variant
(e.g. divertimento SINONIMO_N_DE alegria)
```

Figura 2.1: Caracterização semântica de palavras em PAPEL [24].

por várias equipas de diferentes países são postos à prova e comparados os resultados do sistema na tarefa de análise de sentimentos e na deteção de mensagens relevantes (que referem alguma entidade em particular). O RepLab 2013 focou-se na monitorização da reputação de entidades em mensagens do Twitter. Essas entidades poderiam ser companhias, organizações ou celebridades e o sistema deveria ser capaz de identificar nas mensagens essas referências, catalogá-las de acordo com o seu tópico e retornar uma ordenação baseada na importância que cada uma das mensagens poderia implicar para uma entidade e sua reputação.

2.3 Classificação de palavras

A maioria dos trabalhos que utilizam o método sem supervisão, utilizam a classificação de palavras através de dicionários sintáticos do tipo WordNet[2] ou PAPEL[3], em Português que tem sido utilizado também em diversos trabalhos[24]. Estes dicionários têm como objectivo classificar cada palavra quanto à sua categoria ou *Part-Of-Speech (POS)*, como adjectivos, nomes ou verbos. Outro uso destes dicionários é encontrar sinónimos para as palavras de forma a ser menos complexa a análise e classificação das mesmas e encontrar as raízes de palavras, muito útil quando estas se apresentam, por exemplo, sob a forma de diminutivos.

Estes dicionários podem ser complementados entre si. De forma a alargar a abrangência da lista de sinónimos de palavras presentes no PAPEL em Silva et al[28], foram adicionados também os dicionários TeP[4] e DicSin[5] que são dicionários de Português do Brasil. O conjunto destes 3 dicionários totalizaram 87.327 *lemmas* distribuídos em 136.913 pares de sinónimos dos quais 36.326 são adjectivos.

O PAPEL devolve as relações semânticas das palavras representadas em "triplos", através da estrutura exemplificada na figura 2.1.

Os "triplos" devolvem o tipo de relação semântica entre duas palavras, denominadas por *arg1* e *arg2*. Essa relação pode variar entre sinonímia[6] ou hiponímia[7], identificando também a categoria da palavra, ou POS dos argumentos. No caso da figura, o POS de "divertimento" é "N" de nome. O PAPEL tem vindo a ser constantemente actualizado e melhorado. A última versão 3.2, foi lançada a 31 de Outubro de 2012 e contém cerca de

[2] http://wordnet.princeton.edu/
[3] http://www.linguateca.pt/
[4] http://www.nilc.icmc.usp.br/tep2/
[5] http://dicsin.com.br/content/dicsin_lista.php
[6] http://www.priberam.pt/dlpo/default.aspx?pal=sinonímia
[7] http://www.priberam.pt/dlpo/default.aspx?pal=hiponímia

190 mil relações contabilizando cerca de 95 mil palavras diferentes.

2.3.1 Polaridade de palavras

Um dos objectivos dos trabalhos feitos nesta área, foca-se em classificar a polaridade dos textos de opinião dos utilizadores como positivos ou negativos. Como vai ser descrito, este seria um problema bastante fácil de resolver caso palavras como "ruim" e "não" tivessem sempre uma polaridade negativa e palavras como "bom", "excelente" ou "gostei" fossem sempre palavras que demonstrassem um sentimento positivo. No entanto, não é assim tão fácil. As palavras nem sempre têm o mesmo significado e dependendo do contexto uma palavra positiva pode estar integrada num contexto negativo e vice-versa. Na frase, "ver este filme não é perder tempo" existem duas palavras que podem significar um contexto negativo, as palavras "não" e "perder". Analisando superficialmente podia-se dizer que o sentimento da frase é negativo dado que contém duas palavras com polaridade negativa e nenhuma positiva, no entanto através da percepção humana entende-se perfeitamente que o sentido da frase é positivo e demonstra uma crítica positiva a um filme.

Para além do contexto, também o tópico da opinião influencia o sentido da frase. Por exemplo a frase[18], "deverias ler o livro", teria certamente uma conotação positiva se se tratasse de uma opinião referente a um livro mas teria uma conotação negativa caso se trate de uma opinião a um filme baseado num livro. Estes são pontos que refletem a dificuldade que se enfrenta quanto à classificação de textos e convém analisar não só as palavras que o constituem mas também outros factores como o contexto e o tópico onde o mesmo se insere.

Pang et al[19] analisou qual seria o resultado com a classificação manual de algumas das palavras mais utilizadas dentro do contexto da opinião. O tema em estudo era a análise a filmes e foram dadas a duas pessoas as tarefas de apresentar algumas palavras chave com conotação negativa e positiva que usariam nessa análise, para descrever o filme num comentário que eventualmente tivessem de fazer. Os voluntários utilizaram palavras como "brilhante", "excelente" e "mau", "ruim" para descrever conotações positivas e negativas respectivamente. Após um teste com estas palavras introduzidas manualmente pelos voluntários foram obtidas as taxas de acerto de 58% e 64% para cada um. Em seguida, foi feita uma análise automática ao *corpus* e dele foram identificadas e extraídas as palavras mais utilizadas em comentários positivos e negativos. Dessa análise, resultaram palavras como "lindo" ou "amor" para identificar palavras com sentimentos positivos e palavras como "mau" ou "estúpido" mas também sinais de pontuação como "?" e "!", que estavam frequentemente em textos com sentimentos negativos. Normalmente, nada levaria a pensar aos voluntários que sinais de pontuação poderiam ser marcas frequentes em comentários com sentimentos negativos. Com o mesmo número de palavras-chave que foram utilizadas pelos voluntários, a taxa de acerto atingida pela classificação de polaridade automática foi de 69%, um valor mais alto que em ambos os manuais.

Classificadores semânticos de polaridade

A classificação de polaridade de palavras é algo que tem o seu próprio dicionário. Os classificadores mais utilizados são o SentiWordNet [8], utilizado em [33], o dicionário AFINN [9] ou SentiSense [10] para conteúdo em inglês ou o SentiLex [11] para língua portuguesa. O SentiLex é utilizado em alguns trabalhos. No entanto, devido à sua escassa abrangência de palavras na sua primeira versão (era composto unicamente por adjectivos) os resultados obtidos rondaram os 66% de precisão[24]. A versão 2.0 contempla cerca de 7000 palavras e já contempla também verbos, nomes e expressões idiomáticas. Um estudo recente[29] decidiu testar a eficácia de dois dicionários de sentimentos em português: o SentiLex (já referido) e o OpLexicon [12]. Ao último foi reconhecida uma precisão superior. No entanto, importa referir que a versão utilizada do SentiLex foi a primeira que apenas continha adjectivos e o Oplexicon foi complementado com adjectivos do SentiLex. Baseados nestes pontos, acreditamos que não é justo atribuir melhor eficácia a qualquer um dos dois. Melhor resultado poderia trazer a junção dos dois dicionários nas suas últimas versões. Esta solução suscita alguma curiosidade pois é algo que, até à data, pensa-se que não tenha sido objecto de análise.

Estes classificadores podem ser utilizados para a globalidade dos domínios. Porém, a sua eficácia pode ser inferior em relação a classificadores elaborados automaticamente para um determinado domínio. Por exemplo, a palavra "grande" pode ter uma polaridade positiva se o domínio for opiniões a quartos de hotel ou polaridade negativa se o domínio se tratar de máquinas fotográficas compactas.

O DAL[13] (dicionário de emoções em linguagem) é um dicionário em inglês também bastante utilizado para classificação de polaridade de palavras. Apesar de bastante abrangente, muitos autores encontraram problemas para classificar muitas das palavras. A solução encontrada foi utilizar o WordNet para encontrar sinónimos de palavras que estivessem no *corpus* e que não estavam presentes no DAL, a essas palavras eram atribuídas as polaridades encontradas para o seu sinónimo[1]. Este dicionário ainda tem outra particularidade que é de classificar as palavras de acordo com a sua suavidade (*pleasantness*), a força (*arousal*) e a capacidade da palavra lembrar uma imagem (*imagery*), numa escala da 1 a 3. Assim, a palavra "afecto" terá um valor alto em *pleasantness*, a palavra "energia" em *arousal* e a palavra "flor" em *imagery*. Estas características não estão relacionadas e podem fornecer uma informação adicional.

[8] http://sentiwordnet.isti.cnr.it/
[9] http://neuro.imm.dtu.dk/wiki/AFINN
[10] http://nlp.uned.es/ jcalbornoz/resources.html
[11] http://dmir.inesc-id.pt/project/SentiLex-PT_02_in_English
[12] http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php
[13] ftp://ftp.perceptmx.com/wdalman.pdf

Classificadores semânticos automáticos

Silva et al[28], propõe a criação de um dicionário de sentimentos cujas polaridades são calculadas através de grafos de sinónimos. Esta solução é apresentada como uma alternativa às anteriores dado que pode ser adaptada a um qualquer domínio. O estudo apresentado teve uma precisão de 87%, sobre o corpus WPT05[3].

2.4 Deteção de entidades

Um dos objetivos do sistema de reputação é detetar a entidade alvo do sentimento que a mensagem transmite. Essa é também uma das tarefas dos sistemas presentes no Replab. No RepLab do ano 2012, uma das abordagens que se revelaram mais eficazes foi a do sistema Daedalus[32]. Este sistema extraía uma lista de palavras-chave e referências que estivessem presentes na página Wikipédia[14] da entidade, na página oficial, os endereços de *email* e *hashtags*. Quando as mensagens continham algumas dessas palavras-chave e referências as mesmas eram considerados como relevantes. Este sistema utiliza o que a maioria faz, utilização de listas de *stopwords*, negações e intensificadores de polaridade, no entanto, quanto às *stopwords*, foram acrescentadas à lista, conjuntos de termos que poderiam conter o nome da entidade mas que não estivessem diretamente ligados com a entidade em si. Uma das atividades mais comuns das entidades é o patrocínio ao desporto, por exemplo "Liga Zon Sagres". No exemplo anterior, são indicadas duas palavras que poderiam ser identificadas como entidades (Zon e Sagres), no entanto esta expressão está a indicar especificamente o nome de uma atividade desportiva, mas neste caso é o nome de uma liga profissional portuguesa de clubes de futebol. Uma mensagem que tivesse esse termo não estaria diretamente ligada a nenhuma das entidades referidas.

2.5 Técnicas utilizadas em texto

Existem alguns mecanismos de forma a afinar o detalhe do texto que é analisado. Um dos mecanismos mais utilizados é a remoção de palavras *stop* (*stopwords*). Estas palavras são muito importantes na construção de uma frase e a sua frequência pode ser elevada num determinado comentário. No entanto, para a análise, são palavras que não acrescentam valor. Por exemplo, os artigos ou preposições "o", "ele", "para" ou "um" são palavras *stop* e para análise de texto são palavras menos importantes do que verbos, adjectivos ou substantivos. Porém, a remoção deste tipo de palavras deve ser efectuado com cuidado. Em [5] foram encontrados alguns problemas quando se removeram todas as palavras *stop*, adoptando posteriormente a técnica de só as remover se se encontrarem duas palavras *stop* no mesmo *bigram* (conjunto de duas palavras). Podem ser encontradas listas de palavras

[14] http://www.wikipedia.org/

stop já criadas manualmente tanto em Português[15] como em Inglês[16].

Uma das técnicas mais utilizadas no pré processamento e análise das mensagens é a deteção das negações, como utilizado no sistema UIOWA[33]. As negações são palavras como "não" (*not*), "sem" (*without*) e a presença de uma palavra deste tipo em uma mensagem pode ser responsável por inverter totalmente o sentimento geral da mesma. Analisando as mensagens e respeitando sintaticamente a sua disposição, o sentido geral da mesma pode ser o oposto em relação a quando se analisa palavra a palavra unicamente. Por exemplo, atendendo à palavra "gosto", esta é uma palavra que à primeira vista, inserida num comentário, terá sempre uma orientação positiva. Contudo, antes de um verbo pode existir uma palavra que denote uma negação como a palavra "não", tal como um adjectivo que teria uma conotação positiva terá de passar a ter um sentido negativo pois passará a "não gosto". Pang et al[19] e adiciona a todas as palavras a seguir a uma negação a etiqueta *"NOT_"*. Desta forma, as palavras a seguir a uma negação estarão identificadas e passarão a ter outro significado.

Outra das técnicas é a verificação de intensificadores de polaridade. As palavras que podem intensificar a polaridade são palavras como "mais" ou "muito". A presença de uma palavra destas tem o objetivo de reforçar ainda mais uma palavra com um grau de polaridade à posterior. Por exemplo, o conjunto de palavras "mais eficaz" tem um maior grau de positividade em relação a apenas à palavra "eficaz". Este método também é utilizado em muitos dos artigos relacionados como em [33]. O sistema OPTAH[2] acrescenta a isso a deteção de pontuação repetida, por exemplo "!!!" ou letras repetidas, "Nãoooooo", para reforçar uma polaridade.

2.6 Diferentes métodos de análise

A resolução de trabalhos nesta matéria utilizam maioritariamente dois métodos para análise do texto: o método baseado na análise sintática das palavras (métodos sem supervisão) ou através de técnicas de aprendizagem automática (métodos supervisionados).

2.6.1 Método baseado em regras

O método baseado na análise sintática calcula a orientação do texto baseado na soma das polaridades das palavras que dele fazem parte. As polaridades das palavras normalmente são obtidas através de um dicionário de sentimentos, como o já referido SentiLex ou o SentiWordnet. Os dicionários podem ser manualmente criados ou expandidos através de palavras sementes, como em [31][12][17], para se aproximarem mais do contexto dos textos a serem tratados. Este é um método que não precisa de tempo de aprendizagem. Contudo,

[15] http://snowball.tartarus.org/algorithms/portuguese/stop.txt
[16] http://snowball.tartarus.org/algorithms/english/stop.txt
[17] Conjunto de palavras base que têm como função auxiliar na procura de novas palavras para o contexto.

$$\mathrm{SO}(\textit{phrase}) = \log_2 \left(\frac{\text{hits(phrase NEAR “excellent”) hits(“poor”)}}{\text{hits(phrase NEAR “poor”) hits(“excellent”)}} \right)$$

Figura 2.2: Formula de [5] para calcular a orientação semântica de uma frase.

a sua eficácia pode ser comprometida porque o contexto de uma palavra pode variar consoante o domínio onde a mesma é inserida.

Chaovalit et al[5] faz uso da pesquisa por termos no Google[18] para definir a orientação semântica de um comentário. Depois de uma análise cuidada com o classificador Minipar [19], baseado em POS, são selecionados do comentário todos os fragmentos (ou frases) de duas palavras cada, em que, pelo menos uma, seja um adjectivo ou advérbio. A esses fragmentos são calculadas as orientações semânticas de cada um através da fórmula da Figura 2.2 que relaciona os *hits* (número de ocorrências de páginas numa pesquisa no Google) da frase relacionada, juntamente com as palavras *“excellent”* e *“poor”*. Caso uma palavra esteja mais relacionada com a palavra *“excellent”* ser-lhe-à dada uma orientação positiva, caso esteja mais relacionada com a palavra *“poor”* a frase terá uma orientação negativa. A orientação de todo o comentário será a média das orientações de todas as frases que a compõem.

Taboada et al[30] através do sistema SO-CAL, foi conseguida uma média de taxa de acerto de cerca de 79%, entre corpus de vários domínios. Este sistema não só determinava a polaridade de uma frase como também o seu grau de aceitação ou negação.

No RepLab 2012, o sistema UIOWA [33], a polaridade de cada mensagem é determinada apenas através da soma dos termos positivos e negativos que tivessem a máxima polaridade ao invés da soma de todas as polaridades como acontece na maior parte dos sistemas. A melhor precisão foi conseguida através da utilização do dicionário de sentimentos SentiWordNet com uma precisão de 45%, abrangência 34% e medida F de 34%.

2.6.2 Método de aprendizagem automática

O método supervisionado utiliza mecanismos de aprendizagem automática. As técnicas de aprendizagem automática têm como objectivo estabelecer um modelo de classificação através de um conjunto de dados que represente a informação a ser alvo de análise. Esse conjunto de dados é denominado *corpus* e normalmente, o conjunto de informação utilizado para a aprendizagem é do mesmo modelo da informação que servirá para os testes. É uma boa prática não utilizar o mesmo conjunto de dados para as duas tarefas sob pena de influenciar tendencialmente o resultado final. Este tipo de abordagem pode ser mais eficaz porque é aperfeiçoado ao tipo dos dados que serão objecto de análise e domínio

[18] http://www.google.com
[19] http://webdocs.cs.ualberta.ca/ lindek/minipar.htm

em causa. No entanto, a qualidade da aprendizagem depende da quantidade de dados de treino. Uma quantidade elevada de dados de treino pode significar uma melhoria na eficácia mas necessitar de um elevado tempo para a aprendizagem do modelo[5].

Em Pang et al[19] são analisados três dos métodos mais utilizados em aprendizagem automática: *Naive Bayes*, *Support Vector Machines* e *MaxEnt* (máxima entropia). As definições utilizadas foram baseadas em *unigram* (conjuntos únicos de palavras), adicionando às palavras que estejam próximas a uma palavra negativa, a etiqueta *"NOT_"*. Desta forma, o sentido da palavra é o inverso. Este método verificou uma maior taxa de acerto a nível geral. Em *Naive Bayes* a melhor taxa de acerto registada foi de 81.5% e foi obtida utilizando *unigrams* adicionando a etiqueta POS às palavras, de facto este método foi o único que registou uma melhoria com a utilização de etiquetagem das palavras. Em *MaxEnt* o melhor resultado foi com recurso a *unigrams*, registando cerca 81% de taxa de acerto. No entanto, o treino deste método utilizando apenas *unigrams* é muito pesado e decidiu-se limitar a um conjunto de 2633 amostras. No cômputo geral, tanto para *unigrams*, *bigrams*, ou *unigrams+POS* o método utilizando *SVM* registou quase sempre a melhor taxa de acerto e através de *unigrams*, obtendo cerca de 83% de classificações acertadas, o que se pode considerar um resultado bastante satisfatório. De facto, os métodos que retornam melhores resultados são normalmente com a utilização de *unigrams* [19][23].

Em Chauvalit et al [5] são comparados os desempenhos dos dois tipos de abordagem: o modelo baseado em regras e o modelo de aprendizagem automática. Os resultados mostraram uma melhor taxa de acerto utilizando técnicas de aprendizagem automática, com 85% , contra 77%, obtido com métodos de classificação através de modelos baseados em regras. O corpus utilizado foi baseado em críticas a filmes, cujas instâncias estavam catalogadas em duas classes: positivas e negativas. Turney [31], utilizando métodos de aprendizagem automática, conseguiu cerca de 66% de precisão do sistema para o mesmo tipo de corpus.

Um dos melhores sistemas presentes no RepLab 2012 foi o sistema Uned [6] em que o melhor classificador teve uma precisão de 49%, abrangência de 33% e medida F de 31%, com recurso a árvores de decisão (*Random Forest*). Não havia um classificador de aprendizagem automática que demonstrasse uma eficácia superior em todos os sistemas. As melhores eficácias foram atingidas através dos classificadores mais utilizados: SVM, redes bayesianas ou árvores de decisão.

2.7 Sistemas presentes no RepLab2013

Algumas das ideias e abordagens utilizadas no desenvolvimento deste sistema, têm por base as contribuições dos sistemas presentes no RepLab 2013 [20]. Desta forma, considerou-se importante a elaboração de uma secção dedicada a apresentar algumas ideias e conceitos apresentados em sistemas nesta conferência.

[20] http://www.limosine-project.eu/events/replab2013

O objetivo do RepLab 2013, tal como na versão anterior, focava-se na monitorização da reputação de entidades a partir de mensagens obtidas na rede social Twitter. O desafio principal estava dividido em 4 sub-tarefas:

- Filtragem - O sistema deveria ser capaz de identificar se uma mensagem estava, ou não, relacionada com uma entidade;
- Polaridade - Detetar o sentimento (positivo, negativo ou neutro) de uma mensagem e as suas implicações para a entidade;
- Tópicos - Agrupar as mensagens de acordo com o seu assunto ou evento;
- Prioridade - Atribuir uma importância à mensagem em virtude da sua afetação à entidade.

No conjunto de dados estavam referenciadas 61 entidades de 4 domínios: indústria automóvel, bancos, universidades e artistas de música.

No geral 5 grupos participaram na tarefa de deteção de tópicos, 11 na tarefa de classificação de polaridade, 14 na tarefa de filtragem e 4 na tarefa de prioridade. Os sistemas apresentados em seguida dizem respeito à tarefa de polaridade.

O sistema *SZTE* [10], na sua versão número 8, foi aquele que obteve melhor precisão com 69%, 48% de cobertura , 34% de abrangência e 38% de Medida F. No pré-processamento eram aplicados alguns dos processos e técnicas de processamento mais utilizadas. Por exemplo, a redução das palavras à sua raíz (lematização), a deteção e atribuição de valores de polaridade em *emoticons*, remoção de carateres estranhos, remoção de carateres repetidos, por exemplo "*hellooo*" passaria a "*hello*" e a normalização de números, de URL, de *usertags* e de sinais de pontuação. Era também aplicada a deteção de termos utilizados em redes sociais e substituídos pela sua forma extensa, por exemplo, o termo "*LOL*" passa a "*laughing out loud*". Estes termos, muitas vezes, exprimem um sentimento e em dicionários comuns os mesmos não são referenciados. Por isso, a substituição era efetuada com recurso a um dicionário específico com uma lista deste tipo de termos mais conhecidos. Assim, já seria possível, através de um dicionário de sentimentos comum, detectar um valor de sentimento a atribuir.

O classificador de aprendizagem automática utilizado foi o *Maximum Entropy* e foram utilizadas caraterísticas, como, o valor de polaridade com utilização do dicionário de sentimentos SentiWordNet, a identificação da presença de carateres repetidos, a presença de palavras com letras maiúsculas e quantidade de palavras que exprimem uma negação. No melhor sistema foram integradas a deteção de entidades e a distância entre estas e as palavras que exprimem sentimentos.

O sistema *diue* [22], desenvolvido pelo Departamento de Informática da Universidade de Évora foi um dos representantes portugueses a participar na tarefa de classificação de

polaridade. Muitas das suas ideias serviram como base para a elaboração do sistema desenvolvido no âmbito deste trabalho.

Tal como neste trabalho, para o processamento e análise da informação, foi utilizado o pacote de ferramentas NLTK para Python. O processamento da informação começa por separar as palavras através de pontuação ou de espaços em branco. Em seguida aplica-se a lematização através de WordNet. Para a determinação de sentimento o *diue* utilizou 3 léxicos de sentimento: AFINN, SentiWordNet e um léxico utilizado em [15], treinado a partir de um corpus contendo críticas a produtos.

O sistema submetido foi treinado e classificado com recursos ao algoritmo de aprendizagem automática baseado em árvores de decisão presente no NLTK.

O modelo de dados era composto por 18 caraterísticas. As mais relevantes relacionavam a posição da entidade com as palavras que demonstravam sentimento, como por exemplo, as presenças de negação e de termos polarizados, antes e após a entidade. O resultado do sistema submetido foi de 55% de precisão com 25% de Medida F.

O outro sistema português, POPSTAR [9], veio por intermédio do INESC-ID (Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa)[21]. Como os anteriores sistemas, recorre também a uma abordagem de aprendizagem automática com recurso a regressão logística.

O modelo de dados que serviu para treino e testes englobava a presença de palavras (*bag-of-words*) com pesos atribuídos através da abordagem *Delta-TF.IDF*. Esta é uma abordagem desenvolvida em [17] e neste sistema foram relatados melhores resultados em relação a outros métodos de pesos testados. As palavras com maior valor de entropia eram excluídas pois não forneciam um valor discriminatório entre as classes. Para além disso foram utilizadas outras caraterísticas, como o valor geral de polaridade, número de palavras negativas e positivas, número de sinais de pontuação, *emoticons* ou palavras constituídas por maiúsculas. Uma caraterística que melhorou os resultados foi a adição do título da página à mensagem sempre que no texto estivesse presente um URL. Desta forma, a polaridade era também calculada, com base no título da ligação presente na mensagem. O POPSTAR obteve 64% de precisão e 37% de Medida F.

2.8 Outras aplicações na vida real

A maior parte dos trabalhos referenciados têm por base a análise à informação referente ao contexto de produtos, serviços ou críticas a filmes. Um dos temas mais preferidos mas que causa mais desafios devido à sua complexidade são os dados de análises a filmes [5] [19] [23]. Esta dificuldade surge porque não existe um modelo específico de comentários para um determinado filme. Grande parte das vezes, informações factuais sobre o filme aparecem misturadas com as críticas propriamente ditas e, dessa forma, torna-se compli-

[21] http://www.inesc-id.pt/

cado distinguir o que é uma informação factual que surge no filme de uma análise. Outro problema surge no elevado recurso à ironia. Quando este estilo é muito utilizado torna-se difícil através de técnicas normais de análise, determinar a polaridade de um comentário. Como resultado, os desempenhos obtidos quando se analisaram comentários a filmes não foram tão animadoras quanto a outros tipos de domínios.

A análise a texto proveniente de redes sociais é um dos temas cada vez mais a abordar devido à crescente popularidade das mesmas[29] e essa informação pode ser muito útil para fazer análises em contextos sociais, políticos e económicos.

Em Bollen et al[4] dados provenientes da rede social Twitter foram utilizados para a previsão do mercado de acções.

Em Ratkiewicz *et al* [21] foram comparados dados da rede social Twitter para determinar o abuso de organizações ou pessoas com múltiplas contas no serviço com o objectivo de criar a ilusão de um apoio massivo em relação a um candidato político numa altura de eleições.

Em Duan *et al* [7] foi criado um sistema que detetam os vendedores ou compradores de um sistema *e-commerce* que manipulam ou adoptam maneiras estratégicas ilegais de forma a ganharem pontos de reputação.

2.9 Síntese

A análise de sentimentos em bases de texto é uma das áreas que tem vindo a ganhar maior interesse ano após ano. Uma das causas é a grande utilização de sistemas de partilha de mensagens entre utilizadores. Entre estes sistemas estão as redes sociais, sistemas de comentários a filmes, produtos ou serviços. Estas ofertas podem ser de grande utilidade para as empresas para obterem "*feedback*" relativamente a produtos ou serviços que fornecem.

Neste capítulo são apresentados os diferentes tipos de abordagem de classificação das mensagens. Os resultados divergem consoante as técnicas utilizadas nos sistemas. Regra geral são utilizados dicionários de sentimentos, como por exemplo o SentiWordnet ou o AFINN. A abordagem que na maior parte das vezes retorna melhor taxa de acerto, comparativamente com o apresentado, é com recurso a método de aprendizagem automática. Estas serão algumas das abordagens a desenvolver e testar no sistema desenvolvido no âmbito deste trabalho.

Os trabalhos desenvolvidos e apresentados na conferência RepLab 2013 utilizam as técnicas mais recentes e com melhores resultados nas áreas de análise de sentimentos e reputação da entidade. Estes trabalhos focam-se sobretudo na identificação de entidades e respetiva afetação de sentimentos que são duas funcionalidades a implementar. Desta forma o desenvolvimento e finalidade do sistema apresentado será baseado em algumas das ideias e abordagens utilizadas nestes sistemas.

Capítulo 3

Ferramentas utilizadas

Neste capítulo vão ser descritas algumas das ferramentas utilizadas na elaboração do sistema.

3.1 Natural Language Toolkit - NLTK

O NLTK é uma ferramenta para Python desenvolvida com o objetivo de facilitar e tornar o trabalho com linguagem natural mais eficaz. O seu pacote é composto por variados recursos lexicais e bibliotecas processamento de texto que permitem classificar e identificar as principais caraterísticas em textos. Algumas destas ferramentas foram utilizadas neste trabalho, como as seguintes referidas.

3.1.1 Categorização gramatical

A abordagem utilizada para a classificação de palavras é o módulo `pos_tag` do NLTK. Este é um classificador treinado através do algoritmo *Maximum Entropy* com recurso ao corpus Treebank [1]. Este classificador está treinado para textos em Inglês. Na Tabela 3.1[2] estão referenciadas algumas das categorias gramaticais possíveis pelo sistema de categorização de palavras.

[1] http://www.cis.upenn.edu/~treebank/
[2] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Etiqueta	**Categoria**
CC	Conjunção coordenativa
DT	Determinante
IN	Preposições
JJ	Adjetivo
JJR	Adjetivo, comparativo
JJS	Adjetivo, superlativo
NN	Nome
NNS	Nome, plural
NNP	Nome próprio
NNPS	Nome próprio, plural
PRP	Pronome pessoal
PRP$	Pronome possessivo
RB	Advérbio
RBR	Advérbio, comparativo
RBS	Advérbio, superlativo
VB	Verbo, forma normal
VBD	Verbo, pretérito perfeito
VBG	Verbo, gerúndio
VBN	Verbo, particípio passado
VBP	Verbo, presente singular
VBZ	Verbo, 3ª pessoa presente singular

Tabela 3.1: Lista de categorias de palavras (POS) mais frequentes e importantes neste sistema.

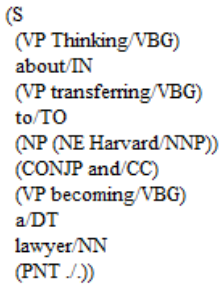

Figura 3.1: Deteção de entidades em uma mensagem.

3.1.2 Detetor de entidades

O método de deteção de entidades utilizado é o método `NE_Chunk` do NLTK, treinado com o algoritmo de aprendizagem automática *Maximum Entropy* através do corpus *Automatic Content Extraction* (ACE) [3] desenvolvido pelo *Linguistic Data Consortium*[4].

Esta técnica de deteção de entidades funciona através de vários passos. O primeiro passo, muito importante, é a correta categorização POS de todas as palavras presentes na mensagem. Após a definição da categoria POS são feitos conjuntos de palavras que congreguem vários elementos da frase, como nomes, verbos ou determinantes. Para a extração de entidades o sistema analisa os vários conjuntos feitos na árvore sintática que representa a frase.

A Figura 3.1 apresenta o resultado do detetor de entidades a partir de uma mensagem presente no conjunto de dados utilizado neste sistema. Neste exemplo foi encontrada uma entidade *"Harvard"* que é identificada com o nó *NE*. O tipo desta entidade é identificado, corretamente, como um nome próprio e está integrada num bloco de nomes com a identificação *NNP*. Estes blocos podem agrupar um ou mais nomes próprios e várias entidades. A entidade pode ser constituída através de uma ou várias palavras.

3.1.3 Dicionários de sentimentos

Os dicionários de sentimentos são conjuntos de palavras manualmente editadas com as suas respetivas polaridades representadas. A polaridade de cada palavra é, normalmente, representada por valores numéricos e indicam se a palavra tem conotação positiva, negativa ou neutra. Em seguida vão ser apresentados os dois dicionários utilizados no desenvolvimento do sistema.

[3] http://catalog.ldc.upenn.edu/LDC2005T09
[4] https://www.ldc.upenn.edu/

POS	ID	ValorPos	ValorNeg	Termos	Glossário
a	00024996	0	0.125	new#11	unfamiliar; “new experiences”
n	14208438	0.125	0.5	neuropathy#1	any pathology of the peripheral nerves

Tabela 3.2: Exemplos de termos no dicionários de sentimentos SentiWordNet.

SentiWordNet

O SentiWordNet[5] é um dos dicionários mais utilizados e com maior abrangência gramatical. É composto por cerca de 38000 palavras derivadas do léxico WordNet e as suas palavras estão categorizadas conforme a sua gramática. Desta forma os valores de polaridade podem adquirir diferentes valores conforme a categoria gramatical de cada uma das palavras.

As palavras adquirem três métricas: positividade, negatividade e objetividade; e os valores variam entre 0 e 1. O valor de objetividade de uma palavra é calculado através da formula: `Obj = 1 - (Valor Positividade + Valor Negatividade)`.

Com uma complexidade maior, em relação a outros testados, este dicionário oferece um conjunto de informação que útil ao sistema e que a sua aprofundada e correta implementação poderá trazer benefícios para o sistema.

Na Tabela 3.2 estão apresentados dois exemplos de termos caraterizados no dicionário de sentimentos SentiWordNet. O dicionário oferece vários tipos de informação:

- POS - Categoria gramatical;
- ID - Número de identificação (único);
- ValorPos - Valor de positividade;
- ValorNeg - Valor de negatividade;
- Termos - O termo a analisar, com a identificação relativamente ao conjunto de sinónimos onde se insere;
- Glossário - Conjunto de termos que ajudam a definir o domínio onde o termo a analisar se insere.

AFINN

O dicionário AFINN[6] é composto por 2477 palavras manualmente anotadas de -5 a 5 conforme o seu valor de negatividade ou positividade respetivamente. Este é um dicionário

[5] http://sentiwordnet.isti.cnr.it/
[6] http://neuro.imm.dtu.dk/wiki/AFINN

Palavra	Valor
best	3
block	-1
celebrating	3
censor	-2

Tabela 3.3: Exemplos de palavras presentes no dicionário de sentimentos AFINN.

com termos gerais e que não traz mais informação para além dos valores de polaridade. Desta forma é bastante fácil e rápida a sua implementação no sistema.

A Tabela 3.3 apresenta alguns exemplos de palavras e o seu respetivo valor de polaridade presentes no dicionário de sentimentos AFINN.

Capítulo 4

Trabalho desenvolvido

4.1 Sensor de reputação

O objetivo deste sistema não passa apenas por detetar o sentimento geral de uma mensagem com base na polaridade das palavras que a compõem. São desenvolvidas um conjunto de técnicas e abordagens em seguida apresentadas que não só integram a tarefa de deteção de sentimento mas também a sua relação com a entidade que está mencionada.

Quando se analisa uma frase deve-se ter em conta o sentimento demonstrado pelas palavras que a compõem mas também de que maneira essas palavras e sentimentos estão relacionados com a entidade. Desta forma, o sistema deve ser capaz de realizar três tarefas complementares. A primeira prende-se com a capacidade de detetar a entidade. No caso do corpus RepLab, serão apresentadas as entidades a identificar, na Secção 4.2. A deteção do sentimento geral através da polaridade das palavras e a determinação de que forma esse sentimento afeta a reputação da entidade detetada são outras tarefas a realizar. Tendo a capacidade de as realizar, o sistema torna-se um sensor de reputação de entidades.

4.2 Conjunto de dados

O corpus do RepLab 2013 consiste num conjunto de mensagens retiradas da rede social Twitter. O conteúdo destas mensagens poderá ser objetivo quando o seu teor não representa uma opinião, ou subjetivo, quando o conteúdo da mensagem indica um estado de espírito, que pode refletir uma opinião sobre um determinado produto ou entidade. Neste corpus estão representadas 61 entidades, desde empresas da industria automóvel, entidades bancárias, escolas ou artistas. O conteúdo das mensagens não foi alterado mantendo-se,

	Número de mensagens
Positivo	13462
Negativo	3164
Neutro	6282
Total	22908

Tabela 4.1: Número de mensagens por cada classe e o seu total.

por isso, a estrutura original das mesmas e as suas caraterísticas principais, como os *tags*, siglas, abreviaturas, ligações ou *emoticons*. O comprimento médio das mensagens presentes no corpus é de 100,82 caracteres.

Após uma análise ao conteúdo do corpus foi detetado que havia algum ruído e que poderia distorcer o resultado do sistema. Posto isto, foram filtradas as mensagens que:

- Entradas em branco - mensagens que tinham sido apagadas pelos seus utilizadores;
- Mensagens repetidas - não seria eficiente estar a analisar duas vezes o mesmo conteúdo;
- Mensagens em diferentes línguas - apenas foram contabilizadas as mensagens escritas em inglês.

Depois de feita uma filtragem das mensagens, a contabilização do conjunto de dados utilizado está descrito na Tabela 4.1. A partir da análise da tabela pode-se conferir que a quantidade de mensagens com conotação positiva é significativamente superior e as mensagens de classe negativa estão muito pouco representadas. Para efeitos de treino e desenvolvimento deste sistema foi dividido o conjunto de dados com uma quantidade de instâncias iguais para treino e teste. Os resultados obtidos no decorrer dos vários testes intermédios estão representados na seguinte secção.

4.3 Técnicas utilizadas

Nesta secção vão ser apresentados alguns resultados intermédios decorrentes do desenvolvimento do sistema. Estes resultados utilizam métricas definidas detalhadamente na secção 5.1.

4.3.1 Pré-processamento do texto

As mensagens por intermédio das redes sociais são muito descritas como difíceis de analisar não só pelo seu tamanho reduzido mas também pelo uso frequente de acrónimos, palavrões, *emoticons*, *Uniform Resource Locator* (URL), *hashtag*, etc. As *hashtags* são palavras muito utilizadas em mensagens nas redes sociais e são precedidas pelo símbolo " # " . Servem para "catalogar" e agrupar mensagens dentro do tópico especificado nesse *hashtag*.

Algumas das caraterísticas podem até trazer alguma informação importante no que à análise diz respeito mas outras podem ser consideradas como ruído na medida em que não apresentam qualquer tipo de informação e como tal devem ser feitas algumas alterações às mensagens originais. As mensagens originais foram alteradas de forma a que seja feita uma análise mais eficaz através dos seguintes métodos:

- Substituição de *emoticons* - Foi utilizado um dicionário de *emoticons* de forma a que fossem substituídos por "happy" ou "sad" cada vez que fosse encontrado numa mensagem. Os *emoticons* foram categorizados por positivos ou negativos, por exemplo: ":)" seria substituído por "happy" e ":(" por "sad". Desta forma já seria possível atribuir um sentimento.

- Remoção de URL - Foram identificados e removidos os endereços para páginas. Para efeitos de análise do texto esta informação não é relevante e poderia causar ruído.

- Tratamento de *hashtags* - Estas são umas das caraterísticas mais comuns em mensagens via Twitter. Foi verificado durante este trabalho que frequentemente estas referências poderiam expressar um sentimento. Desta forma, foi retirado o carater # presente no início de cada uma e a *hashtag* tratada como se de uma palavra normal se tratasse.

- Remoção de *Usertags* - As *Usertags* são referências a outros utilizadores da rede social e por isso não acrescentam valor relevante ao conteúdo.

- Alteração de siglas e termos comuns da *Web* - Termos como *"LOL"* ou *"BRB"* são siglas já bastante conhecidas e utilizadas em qualquer mensagem via Web. Tendo como base uma ideia em [10] foi utilizado um dicionário de siglas e termos mais utilizados. Por exemplo, os termos como *"LOL"* (*Laughing out loud*) ou "BRB" (*Be right back*) são substituídos pelos seus correspondentes significados por extenso. A lista foi retirada de *chatslang.com* [1].

- Remoção de caracteres estranhos - Caracteres como $, %, & ou * são removidos do texto por não oferecerem nenhum tipo de informação relevante.

- Divisão de palavras pelas maiúsculas - É bastante comum em mensagens curtas se encontrarem palavras com algumas letras maiúsculas pelo meio. Essas palavras utilizam-se muito quando são referenciadas *hashtags* ou então apenas para poupar espaço. O sistema procura essas palavras e divide-as por intermédio dos seus caracteres maiúsculos, por exemplo, a palavra *"GreatService"* vai passar a ser duas palavras, *"Great Service"*.

[1] http://www.chatslang.com/terms/common

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.617	0.462	0.528	-
Negativo	0.194	0.350	0.249	-
Neutro	0.288	0.327	0.306	-
Total	-	-	-	0.304

Tabela 4.2: Resultados da primeira abordagem, sem supervisão e através da polaridade dos termos no SentiWordNet.

4.3.2 Classificação da polaridade do sentimento baseada em regras

A primeira experiência para o protótipo inicial teve como base os dicionários de sentimentos AFINN 3.1.3 e SentiWordNet 3.1.3. O primeiro objetivo era testar qual a eficácia de cada um e analisar a abrangência e resultados da junção entre os dois na análise de sentimentos em mensagens.

No caso da junção dos dois dicionários, como as métricas eram diferentes teve de se adotar um método para uniformizar os valores. Sendo assim, optou-se por adaptar o SentiWordNet aos valores do AFINN, juntando todas as palavras com polaridade que não estivessem neste último. Para os valores de polaridade presentes no SentiWordNet maiores de 0.65 o novo valor de polaridade seria de 4, entre 0.50 e 0.625 o novo valor seria 3, entre 0.20 e 0.49 seria de 2 e entre 0 e 0.19 seria de 1. Esta mesma regra também foi adotada para valores negativos ganhando valores inversos. A única razão para estes valores foi encontrar uma semelhança entre os limites em cada um dos dicionários, em que valores maiores no SentiWordNet teriam de obter valores mais elevados de acordo com o AFINN e vice-versa.

A primeira versão consistia apenas em determinar o sentimento geral de uma mensagem com base na polaridade dos termos que a compunham. O valor final era resultado da soma das polaridades de todos os termos encontrados. Os melhores resultados, em termos de acerto global na classificação de polaridade, foram conseguidos considerando uma mensagem como positiva caso a soma das polaridades dos seus termos fosse maior do que 0, negativa caso fosse menor que -2 ou neutra caso o valor obtido estivesse compreendido entre esses dois valores. O único pré-processamento que se efetuou às mensagens foi substituir os *emoticons* por texto que representasse sentimentos semelhantes, por exemplo, ":)" seria substituído por *"happy"*.

Os resultados obtidos na análise das mensagens utilizando os dicionários SentiWordNet, AFINN e a junção dos mesmos, estão representadas nas Tabelas 4.2, 4.3 e 4.4 respetivamente.

Como verificado pelas tabelas referidas, a melhor abordagem é utilizando os dois dicionários de sentimentos em conjunto. Tendo isso em consideração, esta abordagem e estes valores vão servir como base para futuros desenvolvimentos com o objetivo de melhorar a eficácia geral do sistema de classificação.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.669	0.415	0.512	-
Negativo	0.398	0.186	0.253	-
Neutro	0.298	0.620	0.402	-
Total	-	-	-	0.327

Tabela 4.3: Resultados da primeira abordagem, sem supervisão e através da polaridade dos termos no AFINN.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.637	0.542	0.586	-
Negativo	0.304	0.236	0.265	-
Neutro	0.291	0.416	0.342	-
Total	-	-	-	0.346

Tabela 4.4: Resultados da primeira abordagem, sem supervisão e através da polaridade dos termos no conjunto dos dicionários AFINN e SentiWordNet.

A experiência seguinte foi a lematização de palavras, seguindo a metodologia do trabalho [32]. Para isso, foi utilizado o recurso WordNet. Os valores estão apresentados na Tabela 4.5 e, tal como demonstrado em outros trabalhos recentes, como em [32], pode-se verificar um ligeiro aumento na taxa de acerto. Este aumento foi conseguido com a melhoria da cobertura e medida F na categoria dos positivos.

A deteção da negação, foi a seguinte alteração a implementar. Com um conjunto de palavras como "not", "never" ou "neither" numa mensagem a polaridade de uma palavra conseguinte já não terá o mesmo valor. Na maioria das vezes, a presença de uma palavra deste tipo não significa que a polaridade da expressão seguinte seja totalmente invertida, como demonstrado em [30]. Ao invés disso, o método adotado foi retirar menos valor a expressões com maior polaridade quer negativa quer positiva. Apenas uma expressão com um baixo valor de negatividade ou positividade poderá reverter a polaridade. A palavra "excellent" com um grau de positividade no valor de 4 confrontado com uma negação antes, do tipo "not excellent" não indica uma insatisfação mas sim um decréscimo na satisfação logo em vez de 4, o valor de positividade ficará apenas em 2. O método adotado foi retirar ou adicionar 2 valores sempre que o sentimento de a palavra a seguir à negação fosse maior do que 2 ou menor que -2 e inverter caso fosse compreendida entre -2 a 2. A taxa de acerto obtida com a introdução deste método aumentou para 35,4%.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.636	0.564	0.598	-
Negativo	0.298	0.228	0.259	-
Neutro	0.295	0.402	0.341	-
Total	-	-	-	0.352

Tabela 4.5: Resultados da atribuição de polaridade com o conjunto de dicionários aplicando a lematização.

Após remoção de *hashtags* verificou-se uma redução na eficácia do sistema, de onde se conclui que certos *hashtags* são utilizados para representar emoções ou opiniões.

Afetação à entidade alvo

Com base nas metodologias de deteção de entidades e àrvores sintáticas, foi testado o desempenho de uma nova abordagem baseada em regras. Esta abordagem segue a seguinte estrutura de funcionamento:

1. São classificadas as categorias gramaticais (POS) de cada palavra presente na mensagem;
2. Feita a seleção de blocos com conjuntos de categorias de palavras;
3. Detetadas as entidades;
4. Afetação de entidades com outros blocos de palavras possíveis de trasmitirem sentimentos (verbos, advérbios e adjetivos);
5. Cálculo de polaridade do sentimento em relação à entidade;
6. Atribuição da classe da mensagem, definida a partir do valor de polaridade calculado (Positivo, Negativo ou Neutro).

Como referido em 4.3.3, apenas as palavras que forem catalogadas como verbos, adjetivos ou advérbios são consideradas como possíveis fontes de sentimento ou opinião sobre a reputação de uma entidade. Como tal serão apenas essas as contabilizadas para esta tarefa. Para a afetação do sentimento e reputação respeitante às entidades presentes na frase, procedeu-se à análise sintática da frase para determinar possíveis divisões através da pontuação presente. Por exemplo, ",", "." ou "!" são os sinais de pontuação que serviriam para "dividir" a frase em diferentes blocos. No caso de haver apenas uma entidade detetada, todas as palavras que podem transmitir sentimento ou afetar a reputação, de acordo com o considerado neste trabalho, foram afetas a essa entidade, independentemente da sua disposição nos vários blocos da frase. Caso estejam presentes várias entidades e no meio da frase existir algum sinal de pontuação, a relação entre as palavras com sentimento e as entidades seria por proximidade e disposição nesses conjuntos de frases. Na frase *" O meu telemóvel da marca XPTO é bastante rápido, já o meu antigo da marca ABC era muito lento."*. Tal como mencionado, nesta frase encontram-se duas entidades e dois adjetivos. Então, a afetação faz-se consoante a teórica "divisão" da frase em duas e desta forma o adjetivo presente em cada uma das partes afeta a entidade adjacente. Se as entidades estivessem presentes na mesma parte após a divisão por pontuação, as duas entidades eram afetadas pelas palavras possíveis de transmitir sentimento ou afetar a reputação que estivessem contidas no mesmo bloco.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.681	0.208	0.318	-
Negativo	0.366	0.196	0.256	-
Neutro	0.285	0.775	0.416	-
Total	-	-	-	0.362

Tabela 4.6: Resultados da classificação de polaridade e deteção de entidade alvo.

As regras de classificação são efetuadas através da soma da polaridade das palavras afetas à entidade. Caso o valor seja maior do que 0 a classe será definida como Positivo, se for menos do que 0 será Negativo e se for igual a 0 a classe será Neutro. Para testar esta abordagem, uma forma de comparar os resultados foi de apenas contabilizar as entidades previstas para cada uma das frases através das anotações do corpus RepLab. Foram também contabilizadas as entidades detetadas que fossem compostas por várias palavras em que uma delas fosse a prevista, como por exemplo, "Ford Focus" para a entidade "Focus". Todas as outras entidades que foram identificadas para além destas não foram consideradas.

Os resultados obtidos por esta abordagem são apresentados na Tabela 4.6. A taxa de acerto não é muito elevada comparativamente a resultados já apresentados. No entanto, esse valor deve-se também à eficácia da deteção de entidades que é uma tarefa de grande complexidade dentro deste tipo de dados. O resultado final desta abordagem está dependente de dois fatores importantes; eficácia do sistema de deteção de entidades e também do classificador de categorias gramaticais das palavras (POS). O resultado final obtido vai ser sempre condicionado pelo desempenho dessas mesmas tarefas. Embora, como já referido, o resultado não seja tão elevado, esta abordagem tem margem para melhorar quer através de testes com outra abordagem na deteção de entidades com outro sistema de categorização de categorias gramaticais, quer através de um outro conjunto de regras de afetação de sentimentos a entidades.

4.3.3 Aprendizagem Automática Supervisionada

Para os métodos baseados em Aprendizagem Automática, aplicaram-se as mesmas regras de pré-processamento que foram utilizadas nos métodos anteriores. Estes métodos foram exclusivamente treinados e testados através da ferramenta NLTK para Python. Os diferentes classificadores de aprendizagem automática testados foram:

- Naive Bayes
- Árvores de Decisão
- SVM

Todos os classificadores utilizados, tanto na fase de treino como na fase de testes, foram executados com os seus parâmetros normais.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.719	0.782	0.7492	-
Negativo	0.5126	0.4592	0.4844	-
Neutro	0.525	0.4546	0.4873	-
Total	-	-	-	0.647

Tabela 4.7: Resultados de BoW com 600 caraterísticas através do classificador Naive Bayes.

Modelo Bag-of-Words

Com o objetivo de comparar diferentes conjuntos e quantidades de caraterísticas[2] utilizadas na aprendizagem automática dos diferentes modelos, foram testadas várias abordagens:

- *Bag-of-words*
- Lematização de palavras
- Excluíndo *stopwords*
- Valor da polaridade (NB, DT, SVM)
- Etiqueta POS + Polaridade (NB, DT, SVM)
- Bigramas + Polaridade (NB, DT, SVM)

Pretendeu-se comparar de que forma as alterações no tipo e na quantidade de caraterísticas utilizadas no sistema poderiam influenciar os resultados obtidos. Sendo assim, as avaliações de desempenho foram efetuadas utilizando as caraterísticas mais ocorrentes de 300, 600, 1000 e 1200 palavras apenas ou em complemento com outras caraterísticas.

A primeira abordagem, utilizando um dos métodos mais simplista e muito utilizado em diversos trabalhos, é o chamado *Bag-of-Words* (BoW). Este modelo consiste em utilizar apenas as palavras encontradas no conjunto de mensagens como caraterísticas. No vetor que representa cada frase está apenas indicada a presença de cada um dos termos, através de um valor binário. Este é um método que já se provou, por exemplo em [19], ser mais eficaz em relação à quantificação da ocorrência dos termos.

Na Tabela 4.7 estão apresentados os resultados do sistema obtidos através do modelo BoW, com o classificar Naive Bayes. Para uma primeira abordagem, o resultado obtido pode-se dizer que é satisfatório tendo em conta a eficácia geral do sistema que se situou em 64.7%.

O passo seguinte foi usar o BoW, mas tendo como alteração as palavras já lematizadas. O processo é reduzir cada palavra à sua raíz, o que se traduz numa diminuição de palavras

[2]Caraterística, ou *feature*, é um parâmetro a observar pelo algoritmo de aprendizagem.

	300	600	1000	1200
Naive Bayes	0.6492	0.6597	0.6675	0.6716
Árvores de Decisão	0.613	0.615	-	-
SVM	0.6663	0.6697	-	-

Tabela 4.8: Taxas de acerto globais do sistema utilizando diferentes quantidades de caraterísticas juntamente com valor de polaridade.

diferentes mas com significados iguais, como por exemplo, verbos em diferentes modo. Esta também é uma forma de corrigir alguns erros gramaticais. Outra forma de reduzir o conjunto de caraterísticas foi através da remoção de *stopwords*. As *stopwords* são palavras como "the", "if" ou "and" que não apresentam valor nenhum quanto ao significado de cada frase e permitem reduzir substancialmente o número de caraterísticas e por consequência o tempo de processamento do sistema. Aplicando o método de lematização a eficácia do classificar aumentou para 65.2% e aplicando também a remoção de *stopwords* a eficácia conseguida foi de 65.6%, isto para 600 caraterísticas e classificador Naive Bayes. Aplicando estas alterações conseguiu-se um ganho de 0.9% em relação à abordagem inicial.

O passo seguinte foi juntar o modelo BoW com as alterações atrás testadas com a abordagem feita no método inicial baseado em regras, apresentado no capítulo, 4.3.2. Ao modelo BoW é adicionada uma caraterística com o valor da polaridade global da frase. Até este ponto, o método que demonstrou melhor eficácia foi através da adição de uma caraterística utilizando em conjunto as ferramentas de léxico de sentimentos, SentiWordNet e AFINN. Na Tabela 4.8 estão apresentadas as taxas de acerto globais do sistema obtidas através deste método. O classificador SVM foi o mais eficaz mesmo em pequenas quantidades de caraterísticas. No entanto, não foi possível testar com maiores quantidades por falta de recursos computacionais. Em relação ao método anterior, através do classificador Naive Bayes com 600 palavras, esta nova abordagem veio conseguir uma melhoria de 0.3% na taxa de acerto. Com o mesmo classificador mas com 1200 caraterísticas a taxa de acerto foi de 67.16%, a melhor que se conseguiu até a este ponto. No entanto, o classificador SVM conseguiu 66.97% com metade das caraterísticas.

Foram testados mais dois métodos mas que se revelaram com menor eficácia em relação ao anterior. Utilizando a ferramenta WordNet foi catalogada cada palavra com base na sua etiqueta POS. Por exemplo, a palavra *"car"* ficaria *"car_NN"* e as caraterísticas ficariam assim especificadas. Com este método, pretendia-se diferenciar os variados sentidos que a mesma palavra poderia tomar com base no seu contexto. Outra abordagem foi o teste a *bigramas* (sequência de duas palavras) com a caraterística da soma de polaridade da frase.

Estas abordagens conduziram a taxas de acerto inferiores ao obtido anteriormente com recurso a *unigramas* mais o valor da polaridade da frase, como apresentadas na Tabela 4.8. Com a versão utilizando POS e polaridade, a taxa de acerto global do sistema utilizando o classificador Naive Bayes foi de 64.8%, 64.9%, 62.8% e 60.84% para 1200, 1000, 600 e 300 palavras respetivamente, ou seja, cerca de menos 3% no resultado global. Utilizando o classificador de árvores de decisão, o resultado foi de 60% para 600 caraterísticas e 58.6%

para 300. Para SVM, o resultado obtido foi de cerca de 61.20% para 300 palavras, um resultado muito atrás dos 66.6% verificados na versão anterior com recurso ao modelo simples de BoW com a caraterística do valor da polaridade.

Caraterísticas do modelo de dados

Após a normalização das mensagens procedeu-se a um conjunto de técnicas de forma a ser possível retirar informação através da sua estrutura sintática, tal como a sua polaridade e caraterísticas principais.

O modelo de dados é composto por 24 caraterísticas que representam as diferenças nas mensagens presentes no conjunto de dados.

- Percentagem de palavras neutras
- Percentagem de palavras com polaridade positiva
- Percentagem de palavras com polaridade negativa
- Valor máximo de polaridade negativa
- Valor máximo de polaridade positiva
- Presença de palavra com polaridade positiva
- Presença de palavra com polaridade negativa
- Quantidade de palavras com polaridade positiva
- Quantidade de palavras com polaridade negativa
- Quantidade de palavras neutro
- Soma das polaridades de todas as palavras
- Presença de intensificadores positivos
- Presença de intensificadores negativos
- Quantidade de pontuação na frase
- Presença de ponto de exclamação
- Presença de ponto de interrogação
- Presença de negação
- Presença de entidades
- Presença de palavras com caracteres maiúsculos

- Presença de URL
- Polaridade do título do URL
- Valor entropia de palavras com sentimento positivo
- Valor entropia de palavras com sentimento negativo
- Valor entropia de palavras com sentimento neutro

Para o cálculo da polaridade (positiva, negativa ou neutra) de uma mensagem é utilizado o dicionário de palavras com polaridade AFINN 3.1.3. Através deste dicionário são calculados os valores das caraterísticas que estão relacionadas com a **polaridade**.

Os **intensificadores** são palavras que podem significar um reforço a uma palavra com sentimento, quer positivo ou negativo. Exemplo dessas palavras são "mais" ou "pouco". A presença destas palavras está relacionada com frases com teor subjetivo.

As caraterísticas que identificam as pontuações estão implementadas para identificar variadas situações. A caraterística *quantidade de pontuação* adquire um valor numérico que representa a quantidade de sinais de pontuação, por exemplo:

`!"#\$%&'()*+,-./:;<=>?@[\]^_‘{\}~`.

Palavras como "não" ou "nunca" são palavras que demonstram uma negação e que podem inverter tanto a polaridade de um adjetivo ou verbo com sentimento conseguinte ou até o sentimento geral de uma frase. Foi utilizada uma base de dados com palavras como "not", "never" ou "neither" e a presença de uma destas palavras é identificada. O método adotado foi retirar ou adicionar 2 valores sempre que o sentimento da palavra a seguir à negação fosse maior do que 2 ou menor que -2 e inverter caso fosse compreendida entre -2 a 2 [30].

Na maior parte das vezes, a presença de uma palavra deste tipo não significa que a polaridade da expressão seguinte seja totalmente invertida, como demonstrado em [30]. Ao invés disso, o método adotado foi retirar menos valor a expressões com maior polaridade quer negativa, quer positiva. Por exemplo, a palavra "gosto" tem sentimento positivo, as palavras "não gosto" já têm sentimento negativo.

O método de **deteção de entidades** utilizado é o método `NE_Chunk` do NLTK, treinado com o algoritmo de aprendizagem automática *Maximum Entropy* através do corpus *Automatic Content Extraction* (ACE) [3] desenvolvido pelo *Linguistic Data Consortium*[4]. A deteção de uma entidade na mensagem pode significar a subjetividade da mesma.

As mensagens têm a particularidade de conter ligações *URL* para outras páginas. Estas ligações também podem fornecer informação dado que o conteúdo das mesmas está rela-

[3] http://catalog.ldc.upenn.edu/LDC2005T09
[4] https://www.ldc.upenn.edu/

cionado com o conteúdo da própria mensagem. A abordagem seguida neste sistema foi calcular o **valor da polaridade do título** dessas mesmas páginas, como em [9].

O conceito de entropia pode ter vários significados consoante as suas utilizações. Neste sistema, a entropia é utilizada para medir a distribuição de cada palavra em relação às várias classes de sentimentos. Por exemplo, a palavra "*happy*" tem menor entropia do que a palavra "*the*" pois a sua utilização, normalmente, é feita em frases onde o sentimento global é positivo, ao contrário da última em que pode aparecer diversas vezes em frases nas diferentes classes.

A utilização da entropia neste sistema teve como base o trabalho desenvolvido em [9] e foi calculado utilizando a definição feita por *Shannon* em 1948 [27].

A função de cálculo da entropia utilizada é:

$$\mathrm{H(X)} = \text{-}1 \sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \text{ [27]}$$

O X é a palavra objeto de cálculo; n representa as 3 classes onde a palavra pode estar presente (positiva, negativa e neutra); e $p(x_i)$ a probabilidade de a palavra surgir num determinado texto dentro da classe.

O primeiro passo foi realizar a soma de ocorrências de cada palavra em textos para cada uma das classes e em seguida, para cada uma das classes, foi calculada a entropia de cada uma das diferentes palavras do corpus. A entropia foi calculada com base nas instâncias da classe que incluem a palavra a dividir pelo número total de instâncias dessa classe. A entropia global da palavra foi calculada somando o resultado da entropia nas 3 classes.

O resultado da entropia não é indicativo do grau de sentimento de uma palavra, apenas é dada uma indicação da distribuição da sua ocorrência ao nível das diferentes classes. Dando como exemplo os valores H(one) = 0.114 e H(nicely) = 0.0007. Através dos valores não se consegue distinguir qual das palavras representa melhor uma das classes ou um sentimento. No entanto conclui-se que a palavra "*one*" está mais uniformemente distribuída pelas classes e, por isso, menos relevante em termos de informação que a palavra "*nicely*".

As caraterísticas implementadas com recurso à entropia foram incorporadas através do cálculo das palavras positivas, negativas e neutras de uma frase. Considerou-se a entropia apenas de palavras com valor de entropia menor que 0.004. Este valor foi escolhido após uma análise do valores de palavras mais frequentes. Com este limite tem-se a certeza que as palavras que aparecem com maior frequência são excluídas.

Para cada uma das palavras que estão presentes na mensagem o sistema vai calcular os valores das caraterísticas respeitantes à entropia. Se o valor de entropia da palavra for menor que 0.004, é feita uma pesquisa no conjunto de palavras que ocorrem em mensagens das classes Positivo, Negativo ou Neutro, por esta ordem. Caso ela esteja presente numa frase da classe Positivo é adicionado o respetivo valor de entropia à caraterística *"positive_entropy_value"* e o sistema passa automaticamente para a próxima palavra da frase.

```
perc_neutral+polarity_maxnegative+has_posIntensifier+has_negation+polarity_maxpositive+
perc_negatives+has_positive+polarity_value+has_qtdPunctuation+has_exclamationmark
+polarity_qtdneutral+neutral_entropy_value+has_entities+positive_entropy_value+has_uppercase+
polarity_qtdnegative+has_negIntensifier+polarity_qtdpositive+has_questionmark
+has_url+has_negative+perc_positives+polarity_url+negative_entropy_value
```

Figura 4.1: Lista de caraterísticas utilizadas.

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.8139	0.8605	0.8366	-
Negativo	0.4975	0.6609	0.5677	-
Neutro	0.8342	0.5929	0.6931	-
Total	-	-	-	0.7607

Tabela 4.9: Resultados com todas as caraterísticas e valores de entropia com o classificador Naive Bayes.

Se não estiver presente nas palavras que aparecem na classe das frases com polaridade positiva, é feita uma pesquisa do valor de entropia nas palavras de classe Negativo e, em último caso, na classe Neutro.

A Figura 4.1 apresenta todas as caraterísticas utilizadas para além dos valores de entropia. Como apresentado na Tabela 4.9, o resultado da taxa de acerto para esta abordagem foi de 76.07% utilizando o classificador Naive Bayes. Os valores obtidos de precisão e cobertura para as 3 classes foram bastante positivos, acima de qualquer abordagem anteriormente apresentada.

Seleção de caraterísticas

Nem sempre reunir uma grande quantidade de caraterísticas significa ter um conjunto que abranja todos os resultados possíveis e que garanta uma boa eficácia na classificação do modelo de dados. Na verdade, quantas mais houverem, maior ruído poderá haver na construção de um modelo de dados para a classificação.

O objetivo principal da seleção de caraterísticas mais relevantes consiste em encontrar um subconjunto óptimo de caraterísticas que melhorem, ou pelo menos não piorem, a eficácia do sistema, ao mesmo tempo que reduzem a complexidade do modelo e melhoram a eficiência. Desta forma, o modelo de dados torna-se mais simplificado e menos extenso o que reduzirá o tempo e o custo computacional necessário para o seu treino e classificação. Estes objetivos são conseguidos através das remoções de atributos redundantes (caraterísticas bastante semelhantes em relação a outras presentes) e irrelevantes (que não acrescentam informação importante) para o conjunto de caraterísticas presentes. As diferenças entre estes dois tipos de caraterísticas são bastante exemplificadas em [13].

A abordagem utilizada neste sistema para a seleção de atributos foi com recurso a *Wrapper*. Este é um método muito utilizado e bastante referenciado e é considerada uma ferramenta

poderosa na escolha dos melhores atributos, em [13]. O sistema começa como uma caixa *black-box*, quer isto dizer que não existe conhecimento do algoritmo de classificação utilizado. Na prática o melhor conjunto de atributos é encontrado através de várias iterações com todos os conjuntos possíveis de atributos. Para cada conjunto de caraterísticas é calculada a sua eficácia obtida através do classificador de aprendizagem automática e o conjunto final escolhido será aquele que melhor eficácia retornar.

O sistema de procura para encontrar o melhor conjunto de atributos é baseado na técnica *best-first*. Este é um método que retornou bons resultados em [13].

Para a seleção de caraterísticas foi utilizado o *software* Weka[5]. O Weka é uma ferramenta, desenvolvida em Java, com vários algoritmos de classificação automática para tarefas de mineração de dados. A versão utilizada é a 3.7.7.

Foram efetuados vários testes aos algoritmos de aprendizagem automática já anteriormente testados (Naive Bayes, árvores de decisão e SVM). No Weka os algoritmos utilizados foram Naive Bayes, J48 e LibLinear.

A primeira parte consistiu em carregar o conjunto de dados com todas as caraterísticas descritas na Figura 4.1 para o *software*. No passo seguinte é corrido, para cada um dos classificadores, o método de seleção de caraterísticas através do pacote de seleção de atributos "*ClassifierSubsetEval*" com o método de pesquisa "*best-first*". Para cada classificador, as caraterísticas devolvidas foram diferentes em número e em tipo. O último passo foi correr o sistema com o mesmo modelo de dados mas apenas com as caraterísticas selecionadas como relevantes pelo Weka, para cada um dos classificadores. O objetivo foi comparar a eficácia do modelo original com todas as caraterísticas em relação ao modelo com apenas as caraterísticas relevantes e verificar se, de facto, os resultados seriam melhorados. Os valores que se seguem são originados através do sistema desenvolvido. O Weka apenas é utilizado para encontrar o conjunto de caraterísticas mais importantes para cada algoritmo de classificação automática.

Resultados Naive Bayes A primeira abordagem, com 25 atributos e referenciada na tabela 4.9, retornou uma taxa de acerto já bastante satisfatória (76%). A expetativa era grande para verificar se realmente ainda poderia aumentar o seu resultado positivo. As caraterísticas identificadas pelo Weka, como as mais relevantes, foram apenas 3: *has_positive*, *neutral_entropy_value* e *negative_entropy_value*. Um número bastante reduzido em relação à quantidade original. Neste caso, poderá indicar que existem muitas caraterísticas irrelevantes e/ou redundantes.

Para comparar os resultados foi corrido novamente o sistema com as 3 caraterísticas retornadas pelo Weka. O resultado obtido está apresentado na Tabela 4.10. A eficácia foi melhorada em 6% ficando nos 82.08%. De uma maneira geral todos os valores subiram, apenas a precisão na classe Positivo e a cobertura na classe Negativo tiveram valores um pouco abaixo dos verificados inicialmente na Tabela 4.9. Todos os outros valores foram

[5] http://www.cs.waikato.ac.nz/ml/weka/

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.7835	0.9871	0.8736	-
Negativo	0.9279	0.5584	0.6972	-
Neutro	0.9311	0.591	0.7231	-
Total	-	-	-	0.8208

Tabela 4.10: Resultados após seleção de caraterísticas com o classificador Naive Bayes.

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.6386	0.9326	0.7581	-
Negativo	0.5643	0.1876	0.2816	-
Neutro	0.5897	0.1982	0.2967	-
Total	-	-	-	0.6307

Tabela 4.11: Resultados obtidos pelo sistema com todas as caraterísticas através do classificador de árvores de decisão.

melhorados ou mantidos.

Resultados árvores de decisão

O próximo passo foi testar a eficácia da seleção de atributos para o classificador baseado em árvores de decisão. Os resultados obtidos, com todo o conjunto de atributos, estão apresentados na Tabela 4.11. A taxa de acerto, embora seja inferior à verificada através do classificador Naive Bayes, é positiva. No entanto, existem métricas com valores baixos como é o caso da cobertura para as classes Negativo e Neutro. O objectivo da seleção de atributos passará por não só melhorar a taxa de acerto geral mas também por melhorar essas duas métricas.

Neste caso as caraterísticas selecionadas foram em maior quantidade do que em relação ao anterior algoritmo. Foram devolvidas 16 caraterísticas, nomeadamente: *polarity_maxnegative*, *has_posIntensifier*, *has_negation*, *polarity_maxpositive*, *perc_negatives*, *polarity_value*, *has_qtdPunctuation*, *has_excl*, *polarity_qtdneutral*, *neutral_entropy_value*, *positive_entropy_value*, *has_uppercase*, *has_negIntensifier*, *has_questionmark*, *perc_positives* e *negative_entropy_value*.

Os resultados obtidos no sistema, após a escolha das caraterísticas relevantes, estão apresentados na Tabela 4.12. Em relação à taxa de acerto global, o resultado melhorou cerca de 0.1%, o que se traduz numa taxa de acerto de 63.18%. A nível geral, os valores melhoraram, da mesma forma, apenas residualmente e os baixos valores ao nível da cobertura ficaram iguais.

Resultados SVM

O último classificador a ser testado foi o algoritmo SVM. A Tabela 4.13 apresenta os resultados iniciais com todas as caraterísticas presentes. Dos 3 referenciados, este classificador é aquele que tem menor resultado em termos de taxa de acerto. A cobertura das classes Negativo e Neutro é bastante reduzida e pretende-se aumentar o valor desta medida através da seleção de atributos de forma a ser mais viável a classificação de dados

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.6381	0.9363	0.759	-
Negativo	0.5774	0.193	0.2893	-
Neutro	0.5981	0.191	0.2895	-
Total	-	-	-	0.6318

Tabela 4.12: Resultados após seleção de caraterísticas com o classificador baseado em árvores de decisão.

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.609	0.9771	0.7503	-
Negativo	0.4253	0.1210	0.1884	-
Neutro	0.3824	0.0174	0.0333	-
Total	-	-	-	0.5991

Tabela 4.13: Resultados obtidos pelo sistema com todas as caraterísticas através do classificador SVM.

destas duas classes.

Com este algoritmo foram retornadas 8 caraterísticas relevantes, nomeadamente: *has_posIntensifier*, *has_negation*, *perc_negatives*, *has_positive*, *polarity_value*, *polarity_qtdnegative*, *has_questionmark* e *negative_entropy_value*.

Os valores resultantes do sistema utilizando o classificador SVM com as caraterísticas relevantes selecionadas pelo *Weka* estão apresentados na Tabela 4.14. Em todas as métricas de avaliação houve uma pequena melhoria, mas a baixa cobertura nas classes Negativo e Neutro manteve-se. Desta forma, a classificação de um conjunto de dados, nestas 2 classes, terá uma abrangência e eficácia bastante limitada.

Após a análise dos resultados gerados para os 3 classificadores pode-se concluir que o classificador Naive Bayes com a seleção de caraterísticas relevantes é o que retorna melhores resultados. Uma taxa de acerto de 82%, bem como o conjunto de resultados obtidos na precisão e cobertura para as 3 classes, são resultados bastante bons e que comprovam que o sistema desenvolvido garante uma boa fiabilidade na classificação dos dados em estudo.

Nos 3 classificadores testados foi verificado um aumento geral da taxa de acerto após a seleção das caraterísticas relevantes que foram retornadas pelo *software Weka*. Desta forma, pode-se concluir que a utilização desta ferramenta trouxe uma mais valia no que

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.6112	0.9774	0.7521	-
Negativo	0.4368	0.1222	0.191	-
Neutro	0.3913	0.024	0.0452	-
Total	-	-	-	0.601

Tabela 4.14: Resultados após seleção de caraterísticas com o classificador baseado em SVM.

respeita à fiabilidade, eficácia e eficiência de todo o sistema.

Categorias gramaticais

As categorias gramaticais, também conhecidas e doravante denominadas como *Part-Of-Speech* (POS), servem para identificar a categoria de cada palavra com base na sua definição e contexto dentro de uma frase. A relação de cada palavra com as adjacentes são determinantes para atribuir corretamente a categoria a cada uma das palavras. Um dos grandes problemas encontrados em processamento de linguagem natural é a ambiguidade das palavras. Por exemplo, "rio" e "são" são palavras que podem ter dois significados e caraterísticas diferentes (nomes ou verbos) dependendo do contexto onde estão inseridas. Palavras como as anteriores referidas são denominadas como homónimas. As palavras homónimas são aquelas que se escrevem e lêem da mesma forma mas que têm significados diferentes. Assim, o método a utilizar não pode ser simplesmente atribuir a categoria a uma palavra tendo em conta apenas um dicionário com palavras catalogadas. Este é um processo bastante importante no correto funcionamento dos passos implementados no sistema, a seguir referenciados. Foi utilizado o módulo `pos_tag` do NLTK, mais informação em 3.1.1.

Deteção de entidades

Um dos objetivos do sistema é determinar o sentimento atribuído a uma ou várias entidades referidas nas mensagens. Assim, é importante dotar o sistema de um mecanismo que consiga identificar os diferentes tipos de entidades que podem surgir num bloco de texto como pessoas, empresas, lugares ou organizações.

Esta é uma das tarefas mais estudadas na área de processamento de linguagem natural. Como referido em [26], esta nova abordagem surgiu nos anos de 1990 nas jornadas das *Message Understanding Conferences* (MUC) que decorrem nos Estados Unidos. Ao início o objectivo principal era identificar atividades empresariais ou atividades de defesa através de informação retirada de artigos jornalísticos. Com a evolução, aumento da abrangência dos sistemas e novas necessidades, a identificação das entidades tem vindo a ser alargada para temas como organizações, lugares, pessoas ou até expressões temporais e numéricas. Nos dias de hoje, é muito difícil uma única técnica ser transversal e garanta bons resultados nos mais variados domínios. Uma das causas para esse facto é a complexidade semântica das palavras. Por exemplo, a palavra "*Apple*" pode querer fazer referência à empresa ou ao fruto. A única diferença entre as duas palavras é a maiúscula na letra inicial. No entanto, esta não pode ser considerada uma regra de ouro que garanta de uma forma eficaz a distinção entre entidades e não entidades, especialmente em informação proveniente de redes sociais como é o caso.

O método de deteção de entidades utilizado é o método `NE_Chunk` presente no NLTK, mais informação em 3.1.2.

	Precisão	Cobertura	Medida F	Taxa de Acerto
Relacionados	0.743	0.833	0.785	-
Não relacionados	0.256	0.167	0.202	-
Total	-	-	-	0.662

Tabela 4.15: Resultados do identificador de entidades mencionadas em frases.

O método de identificação de entidades em frases baseia-se sobretudo na correta identificação do POS de cada palavra. A partir desta abordagem, tentou-se implementar um método manual de verificação das entidades. A ideia foi determinar cada nome próprio, com a categoria gramatical "NNP" como sendo uma entidade. A taxa de acerto desta abordagem foi de 77%, uma taxa de acerto superior ao método utilizado e referenciado em seguida. No entanto, a cobertura das mensagens que não tinham qualquer entidade foi de apenas 5%. Isto significa que esta abordagem identificou muitas palavras como entidade onde na realidade não eram. Uma das razões verificadas é que as mensagens obtidas em redes sociais, neste caso via Twitter, têm variadas especificidades como a primeira letra de cada palavra começar com maiúscula ou vice-versa, quando verdadeiras entidades referenciadas não respeitam esse princípio.

A Tabela 4.15 apresenta os resultados da identificação de entidades. A coluna "Relacionadas" diz respeito às frases que continham entidades e por isso são consideradas relevantes. A coluna "Não relacionadas" representa as frases que não continham nenhuma entidade e, por isso, o seu conteúdo não estaria relacionado com nenhuma entidade. Este é um método que pode ser utilizado em qualquer domínio e a sua eficácia é positiva com este tipo de dados. No entanto, o tipo de informação utilizada para treinar o classificador de entidades e o utilizado no sistema (mensagens de redes sociais) é bastante diferente e essa pode ser a causa para não haver uma melhor eficácia neste aspeto.

Árvore sintática

As primeiras abordagens na área da análise de sentimentos centravam-se no cálculo da polaridade de uma frase com base na soma das polaridades de todas as palavras que a constituíam. Os novos métodos utilizados vão mais além e utilizam a análise sintática e morfológica de uma frase para analisar de uma forma mais profunda a forma e relação do sentimento demonstrado na frase com a entidade alvo.

Esta nova abordagem tenta solucionar um dos problemas mais comuns na análise de sentimentos que é a relação com a entidade. Através de um exemplo muito simples é possível verificar pormenorizadamente este problema: "*I like brand X, brand Y is causing problems*". Neste exemplo, estão identificadas duas entidades (X e Y) e estão demonstrados dois sentimentos distintos, um para cada entidade referenciada. Se a análise da frase for feita de uma forma superficial, ou seja, através da polaridade geral, o resultado não traz grande informação para as entidades. Através da análise sintática e morfológica da frase, o objetivo é relacionar o sentimento com cada entidade mencionada. Desta forma,

o sistema conseguirá detetar as entidades que forem referenciadas, através do método anteriormente apresentado na secção 4.3.3, relacionando-as com o sentimento que deriva delas, retornando uma polaridade positiva para a entidade X e menos positiva para Y.

Em [16] é feita uma análise sintática preservando as relações e a ordem das palavras na frase. O método envolve a construção de uma sequência de palavras e a criação de uma árvore de dependências com a representação estruturada da frase. A partir dessas árvores de dependências são extraídos os padrões mais frequentes que representam as relações entre palavras. Os testes apresentados demonstraram grande eficácia e apresentaram melhoramentos comparativamente a abordagens anteriores e, por isso, é uma técnica que deve ser tida em conta em sistemas de análise de informação.

Em [8] é descrito o método *Linguistic Tree Transformation* que serviu como inspiração para a metodologia seguida neste sistema. Este método consistiu essencialmente em reduzir a estrutura das árvores analíticas removendo arestas (palavras) que não fossem catalogadas como nomes, verbos ou adjetivos. Desta forma, a informação mais importante era mantida ao mesmo tempo que se tornava mais eficiente a leitura de cada árvore e reduzia o ruído. Depois disso, foram criados bigramas contendo as relações nome-verbo, nome-adjetivo e verbo-adjetivo que estivessem contidos no mesmo nível hierárquico da árvore.

O método utilizado neste sistema utiliza uma representação da frase como uma árvore horizontal, ou seja, com as folhas de cada árvore ao mesmo nível. A análise sintática da frase é efetuada utilizando uma abordagem desenvolvida em [8]. O conceito da árvore horizontal é criar uma representação da frase através de blocos de palavras. Nestes blocos estão contidas as palavras mais importantes, que podem representar um sentimento como verbos ou adjetivos, entidades ou nomes próprios. Cada bloco pode conter vários tipos de categorias e cada frase pode conter vários blocos de frases.

A Figura 4.2 representa os blocos de categorias de palavras possíveis que serão originados em cada frase. As siglas PNT, CONJP, ADJ, NP, VP representam os blocos com pontuação, conjunções, adjetivos, nomes ou verbos, respetivamente. As duas primeiras classes de blocos servem para identificar possíveis divisões ou relações de dependência entre as palavras da oração. Os blocos ADJ e VP dizem respeito às palavras que podem ser importantes para determinar o sentimento do bloco. Finalmente o bloco NP diz respeito aos blocos com nomes e entidades. Em cada um deles poderão estar integradas algumas palavras de outras categorias como verbos ou adjetivos. Isto pretende representar que as mesmas estão separadas por poucas palavras e que a sua relação é muito próxima.

Através da Figura 4.3 que representa uma frase em que uma entidade e sentimento são descritos, é abordado um exemplo da representação gráfica com base em blocos de palavras e sua árvore analítica. Neste exemplo, a entidade e a palavra com sentimento não estão contidos no mesmo bloco, mas é fiável assumir que o sentimento demonstrado pelo adjetivo "*great*" é atribuído à entidade próxima ("*XCars*") quer pela proximidade, quer por existir apenas uma entidade na frase.

```
grammar = """
          PNT: {<.|,>?}
          CONJP: {<PNT>?<CC>+}
          ADJ: {<RB.*>*<DT>*<JJ.*>}
          NP: {<DT|PRP\$|PP\$>*<NNP.*|NE>+} # Chunk nouns
          VP: {<ADJ>*<RB.*>*<DT>*<VB.*>+<JJ.*|RBR>*} # Chunk verbs
          """
```

Figura 4.2: Código utilizado para criar os nós da árvore analítica com base em categorias gramaticais de palavras.

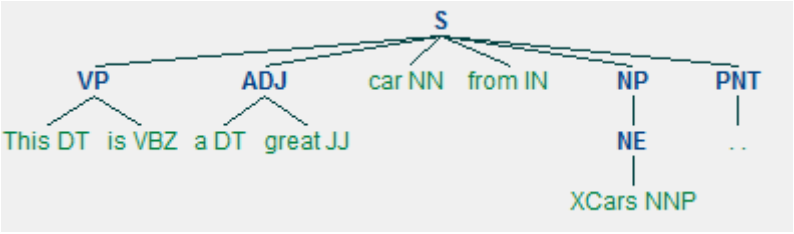

Figura 4.3: Árvore analítica originada a partir de uma frase.

Os blocos representados são aqueles que se consideram os mais importantes para a análise de sentimento, na maior parte das vezes são os verbos e adjetivos que são representativos do sentimento demonstrado. As palavras que não sejam catalogadas dentro destas 3 categorias (verbos, adjetivos e adverbos) e fora dos blocos gerados serão descartadas, uma vez que, muitas vezes não apresentam informação quanto ao sentimento. Para além de não fornecerem informação relevante, a sua integração no modelo de dados pode constituir ruído e retirar eficácia na leitura e classificação do modelo de dados. A sua eliminação também torna o sistema mais eficaz e eficiente.

Tomando como exemplo estas duas frases [19]:

- *"This is a love story!"*
- *"I love this story."*

Apesar das duas conterem quase o mesmo tipo de informação elas são diferentes no que respeita à demonstração de sentimentos. A segunda frase demonstra um sentimento em relação a um nome ou entidade, a palavra "*love*" é na sua forma gramatical, um verbo. A primeira frase é neutra em relação a sentimentos, pois neste caso a palavra "*love*" é um nome e não está a exprimir nenhum sentimento ou opinião sobre determinada matéria. Neste exemplo, é representada a importância que representa a correta categorização de palavras no que respeita à análise de sentimentos.

A Tabela 4.16 apresenta os resultados obtidos através da nova abordagem. Estes valores dizem respeito à classificação por intermédio do algoritmo Naive Bayes com todas as caraterísticas. Analisando a taxa de acerto houve uma melhoria de cerca de 3% em relação à anterior melhor abordagem com caraterísticas semelhantes, referida na Tabela 4.9. A única diferença são as categorias das palavras objeto de análise de sentimento.

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.802	0.917	0.856	-
Negativo	0.625	0.615	0.62	-
Neutro	0.875	0.612	0.720	-
Total	-	-	-	0.792

Tabela 4.16: Resultado do sistema com seleção de polaridade com base em verbos, adjetivos e advérbios.

	Tempo
Total de mensagens	43m48s
Média por mensagem	00,06 seg.

Tabela 4.17: Tempo de execução do sistema.

Com estes resultados conclui-se que, de facto, as categorias gramaticais das palavras devem ser tidas em conta na classificação de sentimentos.

Estes valores são bastante satisfatórios. Uma taxa de acerto elevada em conjunto com os valores altos de precisão, cobertura e Medida F. A classe Positivo, regra geral, foi a que teve melhores resultados nas 3 medidas de desempenho, esse resultado pode estar relacionado com o facto de haverem mais amostras para essa classe. A classe Neutro teve o melhor resultado de todos em termos de precisão dos resultados. A taxa de acerto de 79,2% confere a esta abordagem uma maior confiança em relação à anterior metodologia utilizada.

4.4 Tempo de execução do sistema

Os tempos presentes na Tabela 4.17 dizem respeito ao tempo total despendido pelo programa para classificar o conjunto total das 46511 mensagens de teste. A estes tempos compreendem as tarefas de processamento das mensagens, extração de caraterísticas e classificação das mesmas. O resultado obtido revela um sistema eficiente com cerca de 44 minutos para o conjunto total de mensagens e cerca de 6 centésimas de segundo para analisar uma mensagem. O comprimento médio das mensagens presentes no corpus é de 100,82 caracteres.

Todo o desenvolvimento, bem como os testes, de onde resultam as anteriores medições, foram feitas num sistema com as seguintes caraterísticas:

- Processador - Intel Core i3 (2.13 Ghz)
- RAM - 4 GB (DDR3)
- Sistema Operativo - Microsoft Windows 7 Ultimate (64 bits)

4.5 Utilização do sistema em outras línguas

Este sistema foi desenvolvido e treinado para tratar informação em língua inglesa, mais especificamente sobre mensagens obtidas através da rede social Twitter e como tal não está preparado para tratar eficazmente mensagens escritas em Português.

Para que esta funcionalidade seja implementada neste sistema teriam de ser feitas algumas alterações a vários níveis:

- Dicionários de sentimentos - O dicionário de sentimentos utilizado (AFINN) apenas contem palavras em Inglês. Para determinar a polaridade de palavras terá de ser implementado um dicionário que determine a polaridade em palavras de origem portuguesa como o SentiLex ou o OpLexicon, já referidos;
- Listas de intensificadores e negação - Estas são listas manualmente criadas e no sistema atual apenas foram introduzidas palavras em inglês. Para ser possível a utilização com corpus em Português terão de ser criadas novamente estas listas;
- Lematização - Teria de ser implementada um sistema de lematizaçao que fosse adaptada para a língua portuguesa, por exemplo o sistema Snowball[6] (implementado no NLTK) ou PTStemmer[7];
- Deteção de entidades - Para esta tarefa o sistema atual pode ser mantido, no entanto terá de ser previamente treinado com um corpus em Português. O sistema NLTK dispõe de um conjunto de ferramentas que permitem a fácil integração do sistema em diferentes línguas;
- Categorias POS - Tal como a anterior, esta ferramenta do NLTK pode ser facilmente integrável com outras línguas (neste caso o Português) depois de ser previamente treinado para o efeito.

4.6 Síntese

Como já referido, o conjunto de dados utilizado tem diversas caraterísticas que podem dificultar a sua interpretação e análise. Como tal, foram implementadas diversas técnicas de normalização, apresentadas no capítulo 4.3.1 do texto com vista à aproximação da sua estrutura com a estrutura de textos em linguagem natural de forma a ser mais fácil e eficiente a sua análise por parte do sistema desenvolvido. Apesar de ter sido apenas testado em mensagens provenientes da rede social Twitter, estas mensagens são geralmente difíceis de analisar dada a sua complexidade. A pouca extensão, a diversidade de carateres especiais e referências URL, *hashtags*, entre outras, exigem que seja feito um conjunto de técnicas de processamento do texto. Dada esta complexidade, o sistema desenvolvido

[6] http://snowball.tartarus.org/
[7] http://code.google.com/p/ptstemmer/

adquiriu uma maior versatilidade, prevendo-se que seja capaz de garantir bons resultados a partir de um outro tipo de conjunto de mensagens.

Comparando os resultados pode-se determinar que a abordagem com recurso a aprendizagem automática consegue melhor eficácia que a baseada em regras. O classificador de aprendizagem automática, Naive Bayes, foi o classificador que melhor eficácia e eficiência demonstrou ao longo dos testes efetuados. Este classificador foi o menos exigente em termos de recursos computacionais. Para além de ter conseguido analisar um maior conjunto de dados do que os outros classificadores testados, a precisão obtida foi gradualmente aumentada acompanhando a quantidade de caraterísticas utilizada para a classificação da informação.

Um melhor detetor de entidades poderá significar também uma melhoria na eficácia do sistema, contudo, o resultado obtido através das caraterísticas utilizadas garantem uma boa fiabilidade em detetar o sentimento afeto à entidade encontrada. Dentro das caraterísticas extraídas para a classificação das mensagens destacam-se as que apresentam os valores de entropia das palavras na mensagem. Estas caraterísticas são selecionadas na Secção 4.3.3 , fator que comprova a importância desta abordagem na correta classificação do sentimento e reputação da entidade.

Capítulo 5

Resultados

5.1 Métricas de avaliação

A precisão e a cobertura são duas das métricas mais utilizadas na avaliação de um sistema deste tipo. Tendo como base um sistema de pesquisa de documentos em que alguns são relevantes e outros não relevantes para a pesquisa, a precisão é a percentagem de documentos relevantes dentro dos que foram retornados. Por sua vez, a cobertura é a percentagem de documentos que foram retornados dentro de todos os relevantes no conjunto total dos documentos.

A matriz de confusão é uma tabela utilizada na representação de resultados de aprendizagem automática e tem, como objetivo, a visualização de uma maneira facilitada do desempenho geral de um sistema. Em cada célula da tabela estão contabilizadas as instâncias que foram corretamente e incorretamente classificadas.

A Tabela 5.1 mostra o exemplo de uma matriz. Cada linha da matriz representa a classe real e cada coluna representa a classe prevista da informação. Assim, para saber o total de instâncias que fazem parte da classe Positivo, terão de ser somados os valores previstos que foram corretamente identificados (VP) com os que foram incorretamente classificados mas que pertencem à classe dos positivos (FN).

		Classe prevista pelo sistema	
		Positivo	**Negativo**
Classe real	**Positivo**	VP	FN
	Negativo	FP	VN

Tabela 5.1: Matriz de confusão

As referências nas células têm os seguinte significados:

- Verdadeiros positivos (VP) - Dizem respeito aos documentos que foram corretamente classificados na classe Positivo;
- Falsos positivos (FP) - Os documentos que foram incorretamente classificados como Positivo;
- Verdadeiros negativos (VN) - Os documentos que foram corretamente classificados como Negativo;
- Falsos negativos (FN) - Os documentos que forma incorretamente classificados como Negativo.

Estas métricas servem de referência para os cálculos da precisão e cobertura, apresentadas nas secções posteriores.

5.1.1 Precisão

A precisão do resultado de uma classe é obtida através do número de instâncias que foram corretamente assinaladas nessa classe (VP) sobre todas as instância retornadas, corretas e incorretas (VP+FP).

$\mathrm{P} = \frac{VP}{VP+FP}$

Esta métrica, por si só, não é capaz de retornar grande informação quanto à fiabilidade de um sistema. Uma precisão de 100% indica que todas as instâncias retornadas pertenciam, efetivamente, à classe, mas não indica as instâncias que pertencem à classe e não foram identificadas.

5.1.2 Cobertura

A cobertura é calculada através das instâncias corretamente classificadas(VP) sobre todas as instâncias da classe, encontradas ou não (VP+FN).

$\mathrm{C} = \frac{VP}{VP+FN}$

Tal como na métrica anterior, esta medida não pode ser utilizada isoladamente para determinar a eficácia de um sistema. Por exemplo, se forem retornadas todas as instâncias é possível obter um resultado de 100% não nos fornecendo informação sobre a quantidade de instâncias erradas.

5.1.3 Medida F

A Medida F é uma métrica que combina a precisão e a cobertura. O objetivo é avaliar o resultado de um sistema através das outras métricas referenciadas e combater os problemas já identificados.

Neste trabalho, o resultado obtido é a chamada média harmónica [25] entre a precisão e cobertura. O valor é assim definido na medida em que as duas métricas terão o mesmo peso no cálculo da Medida F.

Medida F = $\frac{2*P*C}{P+C}$

5.2 Avaliação

Para avaliar o desempenho do sistema bem como a sua versatilidade na utilização de diferentes conjuntos de dados foram feitos dois testes. O primeiro incidiu sobre um conjunto de dados semelhante ao anteriormente utilizado no desenvolvimento mas que estava unicamente reservado para os testes finais, ou seja, o sistema não foi treinado com esta nova informação. O segundo teste passou por um diferente tipo de mensagens que foram obtidas através de um sistema de comentários a filmes.

5.2.1 Resultados do sistema com dados de teste

Para o teste e avaliação do sistema foi utilizado um conjunto de 46511 mensagens. Mais uma vez, foram excluídas as mensagens repetidas, em branco e que não fossem escritas em Inglês. Estas fazem parte do corpus RepLab e estavam dedicadas exclusivamente para o teste do sistema, pelo qual nunca foram utilizadas para testar ou treinar o sistema previamente. Foram mantidos os mesmos dados de treino utilizados anteriormente no desenvolvimento do sistema.

As mensagens de teste estão classificadas da seguinte forma:

- Positivo - 27844;
- Negativo - 5925;
- Neutro - 12742.

A Tabela 5.2 representa os resultados obtidos com o corpus de teste através do mesmo sistema que melhor eficácia retornou previamente, como referido na Secção 4.3.3. Estes valores dizem respeito à classificação por intermédio do algoritmo Naive Bayes. Os valores de entropia de cada uma das palavras são os mesmos e os dados para treino são alargados a todo o conjunto que dantes era dividido para fazer as duas tarefas. Desta vez, os dados

	Precisão	Cobertura	Medida F	Taxa de Acerto
Positivo	0.64	0.792	0.708	-
Negativo	0.299	0.247	0.271	-
Neutro	0.359	0.202	0.258	-
Total	-	-	-	0.561

Tabela 5.2: Resultados do sistema com dados de teste utilizando a melhor abordagem com recurso à entropia.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.638	0.798	0.709	-
Negativo	0.292	0.23	0.258	-
Neutro	0.362	0.199	0.257	-
Total	-	-	-	0.562

Tabela 5.3: Resultados do sistema com dados de teste com polaridade afeta à entidade alvo, com recurso ao algoritmo Naive Bayes.

reservados para testes e que nunca foram utilizados para o desenvolvimento do sistema foram utilizados para testar a eficácia da melhor abordagem conseguida com recurso a técnicas de aprendizagem automática e com todas as caraterísticas.

A Tabela 5.3 representa o resultado de outra versão do sistema que é semelhante à anterior mas desta vez os valores de polaridade são calculados como apresentado em 4.3.2. Em vez do valor da polaridade ser calculado com base em todas as palavras possíveis de transmitirem sentimento ou afetarem a reputação de uma entidade, desta vez a polaridade só vai ser calculada com base nas palavras que afetam diretamente a entidade pretendida presente na mensagem. Caso não seja detetada a entidade específica procede-se ao cálculo geral da polaridade. A anterior abordagem, exclusivamente baseada em regras 4.3.2 e com os dados de treino, tem os seus valores apresentados na Tabela 4.6.

Esta abordagem obteve resultados melhores em relação à abordagem semelhante, exclusivamente baseada em regras de afetação de sentimentos e reputação a entidades, demonstrados na Tabela 4.6. A diferença pode ser explicada pela melhor eficácia que é normalmente conseguida através de métodos de aprendizagem automática e pelo recurso à utilização de valores de entropia no conjunto das caraterísticas utilizadas para o treino e classificação do algoritmo utilizado. Relativamente à abordagem anterior, não houve diferença visível na eficácia obtida. A razão para isto pode estar no facto das caraterísticas relativas à entropia terem uma importância maior comparativamente com as relacionadas com a polaridade das palavras.

	Precisão	Cobertura	Medida F	Taxa de acerto
Positivo	0.854	0.72	0.781	-
Negativo	0.75	0.872	0.806	-
Total	-	-	-	0.795

Tabela 5.4: Resultados da classificação de mensagens de críticas a filmes com valores de entropia, por intermédio do algoritmo Naive Bayes.

5.2.2 Resultados do sistema com diferente corpus

Para testar a eficácia do sistema utilizando diferentes conjuntos de dados fez-se uma experiência com mensagens de críticas a filmes. O corpus[1] é constituído apenas por mensagens com polaridades Positivo ou Negativo. Desta forma, foram retirados os valores de entropia para palavras neutras bem como a caraterística respeitante no conjunto de atributos utilizado na classificação com recurso a aprendizagem automática. Este corpus é constituído por 10662 mensagens, com um número igual de mensagens de classe Positivo e Negativo (5331).

O sistema foi treinado e classificado com um subconjunto de dados de quantidade igual, semelhante à abordagem anterior. Como foi tratado um conjunto diferente de dados novos, valores de entropia foram calculados para cada uma das palavras presente no corpus. Esta versão teve como base a melhor abordagem obtida com recurso ao corpus RepLab através de todas as caraterísticas do modelo de dados e valores de entropia indicados na Tabela 4.9. Foi também mantido o valor de entropia máximo de cada palavra (0,004).

O resultado apresentado na Tabela 5.4 é muito satisfatório pois todos os valores obtidos são considerados como bastante fiáveis. Tanto a precisão como a cobertura de cada uma das classes apresentam valores perto ou acima dos 80%. Perante estes valores pode-se afirmar que este sistema, através da abordagem desenvolvida e após o devido treino, é eficaz na tarefa de classificar diferentes tipos de texto.

5.3 Síntese

Com um conjunto diferente de dados para teste, a eficácia retornada foi diferente da obtida em 4.16 com a mesma abordagem. Apesar de ser um valor mais baixo, trata-se de um resultado positivo que comprova que esta abordagem pode ser utilizada mesmo sem um conjunto de caraterísticas, como valores de entropia das palavras, previamente calculadas a partir do mesmo conjunto de dados que vai ser utilizado para os testes.

Os testes efetuados com um corpus diferente revelaram que esta abordagem também pode ser utilizada com um corpus com diferente estrutura. Os corpus compostos por mensagens constituindo críticas a filmes são frequentemente utilizados na área de sistemas de análise de sentimentos. Os valores obtidos pelo sistema utilizando este corpus demonstram um

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data

sistema bastante eficaz na análise do sentimento geral da mensagem. Este corpus pode ser considerado mais fácil de analisar dado que não é necessário considerar a entidade e tratar o sentimento geral da mesma. O valor obtido por este corpus está ao nível ou até melhor em comparação com alguns dos sistemas apresentados no estado da arte que utilizam o mesmo tipo de conjunto de dados.

Capítulo 6

Conclusões

As redes sociais desempenham um papel importante na sociedade moderna. A facilidade de acesso, ausência de custos e abrangência da comunicação, fizeram e continuam a fazer com que as redes sociais se expandam e que ganhem novos adeptos a cada dia que passa. Esta é uma ferramenta não só utilizada para fins lúdicos mas, cada vez mais, também com interesses comerciais.

As marcas, aproveitando-se das potencialidades desta nova forma de comunicação global, estão a adoptar estratégias que as permite estarem em constante promoção da sua marca com os seus clientes. No entanto, com um universo tão grande, tanto de plataformas de redes sociais como de utilizadores, as mensagens geradas diariamente adquirem uma dimensão tão elevada que é impossível retirar algum tipo de informação de uma forma manual.

O objetivo principal deste trabalho foi o desenvolvimento de um sistema de reputação de entidades com base em análise de sentimentos em frases retiradas do *micro-blog* Twitter. Este capítulo aborda as considerações finais onde são apresentadas as principais contribuições e inovações que este trabalho apresentou, o seu resultado prático e algumas ideias de trabalho a implementar numa próxima versão, com vista a melhorar a eficácia do sistema.

6.1 Balanço final

Na fase inicial desta dissertação é descrito e analisado o estado da arte em relação a sistemas da área de análise de sentimentos. É ainda feita uma pequena introdução sobre em que consiste este tópico e as abordagens iniciais que deram origem aos avanços e metodo-

logias mais utilizadas nos dias de hoje. Os trabalhos presentes nos RepLab 2012 e 2013 são frequentemente citados e, como já referido, têm grande influência no desenvolvimento deste sistema devido a serem sistemas que têm objetivos e conjuntos de dados semelhantes.

Foram efetuados vários testes a diferentes abordagens do sistema, nomeadamente a diferentes técnicas de pré-processamento e normalização da informação e classificação das classes das mensagens com recurso a métodos baseados em regras e algoritmos de aprendizagem automática.

Apesar de este trabalho se centrar no desenvolvimento de técnicas de classificação de sentimento em textos através de mecanismos de aprendizagem automática, foram feitos alguns testes utilizando apenas um conjunto de regras. Os resultados desta abordagem estão apresentados na secção 4.3.2 e foram obtidos unicamente através de um conjunto de técnicas de normalização e processamento do texto aplicando posteriormente a classificação do sentimento através de um léxico de sentimentos. Estes resultados são mais baixos em relação aos resultados respeitantes à utilização de técnicas de aprendizagem automática. A razão de serem obtidos resultados mais baixos não quer dizer que esta abordagem seja pior ou melhor, mas que, o sucesso desta abordagem requer um maior tempo para análise da informação e para o desenvolvimento de um conjunto de regras que retornem resultados eficazes. A técnica da lematização de palavras, com 35,2% de taxa de acerto, originou um ganho de eficácia, pelo que se decidiu optar pela sua integração na abordagem a técnicas de aprendizagem automática.

No capítulo 4.3.3 são apresentados os resultados com recurso a técnicas de aprendizagem automática. A classificação com recurso ao valor de entropia das palavras, com os resultados apresentados na Tabela 4.9, demonstrou ser uma caraterística bastante eficaz. O método descrito em 4.3.3, que restringe a classificação do sentimento apenas a alguns tipos de categorias de palavras, provou ser uma abordagem eficiente, na medida em que melhorou a precisão do resultado.

A deteção de entidades era uma funcionalidade tida como um dos objetivos para o sistema. Apesar de a sua eficácia ser um dos pontos menos positivos deste sistema, a sua limitação prende-se com o facto de este tipo de dados ter uma certa complexidade e o detetor de dados utilizado estar treinado para entidades presentes em informação mais estruturada, como por exemplo, as palavras referentes a entidades iniciando em maiúsculas, coisa que não acontece muitas vezes no conjunto de dados utilizado.

O teste final ao sistema devolveu um resultado de cerca de 56% de taxa de acerto. Este é um resultado muito satisfatório tendo em conta que é um sistema que tem duas tarefas complexas inerentes: a deteção de entidades e a afetação da respetiva reputação. Analisando os resultados obtidos no RepLab 2013, que são sistemas que utilizam o mesmo tipo de dados, o resultado obtido coloca este sistema no meio da tabela dos sistemas presentes na tarefa de polaridade. Posto isto, são indicadores que garantem ao autor uma motivação para continuar a aperfeiçoar o sistema através da pesquisa e/ou desenvolvimento de novos métodos e técnicas que garantam melhores resultados.

Através dos resultados obtidos, pode-se concluir que esta abordagem e técnicas utilizadas garantem uma boa fiabilidade em detetar o sentimento geral de uma frase nas 3 classes distintas. O cálculo do valor de entropia de uma palavra relativamente à sua classe demonstra ser uma caraterística com bons resultados na distinção entre as várias classes.

De uma forma geral, pode-se dizer que o objetivo geral deste trabalho foi cumprido. A identificação de sentimentos e sua afetação da reputação são duas tarefas que são concretizadas pelo sistema com uma eficácia satisfatória. Um dos objetivos era implementar uma funcionalidade que trouxesse alguma inovação comparativamente a sistemas relacionados e creio que a entropia foi uma inovação implementada neste sistema com bons resultados.

6.2 Trabalho futuro

Este é um sistema que ainda não está totalmente finalizado e que poderá ser melhorado com a introdução de treinos e testes a partir de um diferente conjunto de dados. Um dos aspetos com mais margem para melhorar é a deteção de entidades. Neste sistema optou-se por utilizar um sistema geral sem nenhum tipo de treino para este conjunto de dados nem para o tipo de entidades tratadas. Este facto gerou uma eficácia mais baixa que pode ser melhorada com um outro tipo de abordagem mais aprofundado tendo por base as abordagens utilizadas em sistemas presentes no RepLab. Algumas das soluções podem passar por utilizar abordagens efetuadas em [20] como a utilização de recursos externos com possível informação relevante sobre a entidade e palavras-chave relacionadas, como a Wikipedia[1] ou Freebase[2] utilizando técnicas de cálculo de similaridade entre a informação obtida e a mensagem alvo de análise.

O dicionário de sentimentos SentiWordNet, apesar de não ter sido utilizado nas versões finais do sistema, é um recurso bastante utilizado, como por exemplo em [10]. Numa futura versão, esta parte também deveria ser mais explorada com a utilização de apenas este dicionário ou feita uma integração mais aprofundada no sentido de integrar os dois dicionários em conjunto. Os valores de positividade, negatividade e objetividade que o SentiWordNet atribui às palavras podem ser úteis para a adição de novas caraterísticas importantes para o modelo de dados para a classificação através de aprendizagem automática.

A adaptação para outras línguas é um dos aspetos já referidos em 4.5. Para além da vantagem óbvia que é ser abrangente, o sistema poderia ganhar em termos de eficiência em ser treinado através de novas grafias resultantes de textos escritos em diferentes línguas.

Outra das implementações poderá ser a correção automática das palavras com o algoritmo da distância de Leveshtein [14], uma das formas utilizadas em motores de busca que existem para eliminar os erros de ortografia. O objetivo será conseguir aproximar as caraterísticas especiais de uma mensagem obtida num *micro-blog*, como o Twitter, com uma mensagem

[1] https://www.wikipedia.org/
[2] http://www.freebase.com/

normal.

Como referido, as caraterísticas que envolveram os valores de entropia das palavras tiveram bastante importância no modelo de dados utilizado na classificação de novas instâncias. Essa importância poderia ser mais explorada, por exemplo, através da implementação de um conjunto de caraterísticas, ou até, um classificador à parte do existente, baseado em BoW. Esse classificador incidiria sobre o conjunto de palavras com maior capacidade discriminatória entre classes, ou seja, com menor entropia.

Referências bibliográficas

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09. Association for Computational Linguistics, 2009.

[2] Alexandra Balahur and Hristo Tanev. Detecting entity-related events and sentiments from tweets using multilingual resources. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[3] David Batista and Mário J. Silva. A statistical study of the wpt05 crawl of the portuguese web.

[4] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 2011.

[5] Pimwadee Chaovalit and Lina Zhou. Movie review mining: a comparison between supervised and unsupervised classification approaches. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04*, HICSS '05. IEEE Computer Society, 2005.

[6] Jorge Carrillo de Albornoz, Irina Chugur, and Enrique Amigó. Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[7] Huiying Duan and Feifei Liu. Building and managing reputation in the environment of chinese e-commerce: a case study on taobao. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12. ACM, 2012.

[8] Brian Eriksson. Sentiment classification of movie reviews using linguistic parsing, 2006.

[9] Joao Filgueiras and Silvio Amir. Popstar at replab 2013: Polarity for reputation classification. In *To appear in: Fourth International Conference of the CLEF initiative*, CLEF 2013, 2013.

[10] Viktor Hangya and Richárd Farkas. Filtering and polarity detection for reputation management on tweets. In *To appear in: Fourth International Conference of the CLEF initiative*, CLEF 2013, 2013.

[11] Ahmad Kamal, Muhammad Abulaish, and Tarique Anwar. Mining feature-opinion pairs and their reliability scores from web opinion sources. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12. ACM, 2012.

[12] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04. Association for Computational Linguistics, 2004.

[13] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, December 1997.

[14] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[15] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.

[16] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 301–311. Springer-Verlag, 2005.

[17] Georgios Paltoglou and Mike Thelwall. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1386–1395, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[18] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, January 2008.

[19] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. 2002.

[20] Arian Pasquali Carlos Soares Jorge Teixeira Fábio Pinto Mohammad Nozari Catarina Félix Pedro Strecht Pedro Saleiro, Luís Rei. Popstar at replab 2013: Name ambiguity resolution on twitter. In *To appear in: Fourth International Conference of the CLEF initiative*, CLEF 2013, 2013.

[21] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and tracking political abuse in social media. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[22] José Saias. In search of reputation assessment: experiences with polarity classification in replab 2013. In *To appear in: Fourth International Conference of the CLEF initiative*, CLEF 2013, 2013.

[23] Franco Salvetti, Christoph Reichenbach, and Stephen Lewis. Opinion Polarity Identification of Movie Reviews. 2006.

[24] AntónioPaulo Santos, HugoGonçalo Oliveira, Carlos Ramos, and NunoC. Marques. The role of language registers in polarity propagation. In Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão, editors, *Computational Processing of the Portuguese Language*, volume 7243 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012.

[25] Yutaka Sasaki. The truth of the f-measure, 2007.

[26] Satoshi Sekine. Named entity: History and future. 2004.

[27] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[28] Mário Silva, Paula Carvalho, and Luís Sarmento. Building a sentiment lexicon for social judgement mining. In Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão, editors, *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2012.

[29] Marlo Souza and Renata Vieira. Sentiment analysis on twitter data for portuguese language. In Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão, editors, *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2012.

[30] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*

[31] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *CoRR*, cs.LG/0212032, 2002.

[32] Julio Villena-Román, Sara Lana-Serrano, Cristina Moreno, Janine García-Morera, and José Carlos González Cristóbal. Daedalus at replab 2012: Polarity classification and filtering on twitter data. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[33] Chao Yang, Sanmitra Bhattacharya, and Padmini Srinivasan. Lexical and machine learning approaches toward online reputation management. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.