Daniel Henriques Rodrigues

Construção Automática de um Dicionário Emocional para o Português

Universidade da Beira Interior

Departamento de Informática

Agosto 2009

Daniel Henriques Rodrigues

Construção Automática de um Dicionário Emocional para o Português

Tese submetida ao Departamento de Informática para o preenchimento dos requisitos para a concessão do grau de Mestre efectuada sob a supervisão do Doutor Gaël Harry Adélio André Dias, Universidade da Beira Interior, Covilhã, Portugal

Universidade da Beira Interior
Departamento de Informática
Agosto 2009

Agradecimentos

Os primeiros agradecimentos vão naturalmente para os meus pais por me terem dado a educação e a formação que tenho hoje. Ao meu irmão e minha cunhada por me terem apoiado e motivado. Dedico esta tese de mestrado a eles porque são as pessoas mais importantes na minha vida.

Finalmente, agradeço a todos os meus amigos, a todo as pessoas da melícia IMPERAT, as pessoas do Hultig e ao Sebastião, que desenvolveu a sua tese de mestrado na mesma área que eu com o mesmos supervisor, e é um grande amigo. E obviamente, agradeço ao Doutor Gaël Dias por me ter apresentado e orientado nesta tese de mestrado.

Obrigado a todos!

Daniel Rodrigues :)

Resumo

Nos últimos anos tem-se verificado um crescente fluxo de conteúdos disponíveis e de fácil acesso na World Wide Web (Web), fazendo com que actualmente haja uma acumulação excessiva de textos de diversas naturezas. Apesar dos aspectos positivos que isto representa e do potencial que acarreta, surge uma nova problemática que consiste na necessidade de desenvolver ferramentas e metodologias capazes de tratar esses mesmos conteúdos ao nível das opiniões e/ou sentimentos neles evidenciados.

A avaliação dos conteúdos Web não é uma tarefa fácil. As técnicas de avaliação dos conteúdos estão inseridos na área de análise de sentimento e muitos são os trabalhos sugeridos. Esta tese segue um rumo diferente, com ela pretende-se avaliar os conteúdos Web para a língua portuguesa europeia. O critério base adoptado é criar as bases para, no futuro, construir classificadores de sentimentos.

Os léxicos emocionais servem de base a grande parte dos métodos que efectuam a análise de sentimento. Apesar de existir uma grande quantidade desses recursos disponíveis para a comunidade científica, depois de muita pesquisa, verificou-se que não existe um recurso semelhante para a língua portuguesa europeia. Com o interesse cada vez maior por parte das empresas ou indivíduos em obter informação sobre os produtos em tempo real a partir dos dados da Web, existe a necessidade de construir um léxico emocional para o português que possa ser utilizado para efectuar a análise de sentimento, para esta língua.

Para colmatar esta falta, construiu-se automaticamente um léxico emocional para o português. Os métodos que efectuam a análise de sentimento utilizam léxicos construídos manualmente ou semiautomaticamente, surgindo o problema do acrescentar conhecimentos linguísticos aos léxicos, inerente ao modo como estes são construídos.

Sendo a identificação dos sentimentos a chave do processo, é necessário saber que os sentimentos são sinónimo de subjectividade. O desafio colocado nesta dissertação é de construir automaticamente um léxico subjectivo para o português europeu, aplicando técnicas estatísticas.

A base da construção do léxico são os corpora que vão ser utilizados. Para efectuar este estudo é necessário um corpus subjectivo (constituído por textos dos blogues) e um corpus objectivo (constituído por textos do corpus jornalístico CETEMPúblico). Os corpora foram escolhidos com base no estudo efectuado por Pais [21] na demonstração da similaridade entre um conjunto de blogues e um corpus jornalístico, relativamente a um corpus constituído por textos subjectivos e objectivos.

Para identificar a subjectividade no texto utilizou-se a informação das categorias morfológicas (part-of-speech) das palavras simples e as palavras compostas (n-grams). Estes indícios de subjectividade foram extraídos com ferramentas que efectuam o processo automaticamente.

Com este trabalho demonstrou-se que é possível construir um léxico de subjectividade para o português europeu, aplicando técnicas estatísticas e utilizando corpora não anotados manualmente e ferramentas para extrair automaticamente os indícios de subejctividade.

Abstract

It has been verified in the last few years, the increasing of data contents and easy access to Web allowing an excessive gathering of text from several natures. Despite of all positive aspects and the promising potential, appears a new problem related to the need of develop tools and methods capable to treat those contents in opinions and feelings.

The evaluation of Web contents is not an easy task. The content evaluation techniques belong to feelings analysis area and there is a lot of proposed works. This thesis follows a different course by evaluating the web contents to European Portuguese language. The initial criterion adapted is to build the bases to create feeling classifiers.

The emotional lexicon is the base to the most of the feeling analysis methods. Regardless of being a lot of available resources to the scientific community, has been verified after big research, that there is no similar resource to European Portuguese language. With the interest of most companies and individuals to gain product information from data web in real time, there is a need to build an emotional Portuguese lexicon that might be use to feeling analysis.

To deal with this missing, an emotional Portuguese lexicon has been built. The feeling analysis methods use lexicons, built manually or semi automatic, appearing the problem of increasing the linguistic lexicon acknowledgment associated to the way of how they are built.

Considering the feelings has the key of the process, it's important to know that feelings are synonymous of subjectivity. The goal of this dissertation is to build an automatic European Portuguese subject lexicon using statistical techniques.

The lexicon construction base is the corpora which will be used. To do this research is required a subjective corpus (built by blog text) and an objective corpus (built by CETEMPúblico journalistic corpus text). The corpora have been chosen based in Pais [21] research in the similarity demonstration between a blog amount and a journalistic corpus, relatively to a corpus built by subjective and objective texts.

To identify the text subjectivity has been used the information from the morphologic categories (part-of-speech) of simple and compound (n-grams) words. These subjectivity evidences were extracted with tools that work with the automatic process.

This work demonstrate that it is possible to build a subjectivity lexicon to the European Portuguese language applying statistical techniques and using manually not noted corpora as well as subjectivity evidences automatic extraction tools.

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

LSA Latent Semantic Analysis

MB Megabytes

ME Mutual Expectation

PMI Pointwise Mutual Information

PLN Processamento de Linguagem Natural

SCL Stuctural Correspondence Learning

SVM Support Vector Machine

UT Unidade Textual

Web World Wide Web

Capítulo 1

Introdução

Com o nascimento da Web nos primeiros anos da década de 90 ocorreram alterações profundas na percepção que se tinha, até então, relativamente aos sistemas de pesquisa de informação[1] em documentos. Estes sistemas passaram a abordar áreas como a modelação, classificação e filtragem de documentos, entre outras. Hoje em dia a Web é uma grande enciclopédia de fácil acesso e ao serviço de todos, que necessita de ferramentas, cada vez mais poderosas e eficientes, para que o conhecimento acumulado possa ser devidamente pesquisado e utilizado. Deste modo urge a necessidade de encontrar novas metodologias para a classificação dos documentos que permitam passar duma visão centrada em dados para uma visão centrada na informação, isto é, conceitos e contextos.

Os sentimentos expressos nos documentos estão relacionados com diferentes tipos de conteúdos. Da análise da definição de subjectividade, verifica-se que estes sentimentos contidos nos textos dos documentos estão directamente relacionados com a subjectividade. Assim, estes sentimentos podem ser classificados em dois conjuntos distintos de classes: objectivo/subjectivo ou positivo/negativo.

Nos últimos anos tem-se verificado um crescente interesse na extracção automática ou semiautomática de opiniões, emoções e sentimentos em documentos que se encontram em suporte digital (análise de sentimento); em fornecer ferramentas e suporte para várias aplicações Processamento de Linguagem Natural (PLN).

[1] Em inglês diz-se Information Retrieval

De um modo geral, a análise de sentimento serve para determinar a atitude de um locutor ou escritor em relação a um dado assunto. Essa atitude pode ser um juízo de valor ou avaliação, o estado emocional do autor ao escrever ou o efeito emocional que o autor pretende ter sobre o leitor.

Na análise de sentimento a tarefa básica consiste em classificar a polaridade de um determinado texto ou frase no documento, isto é, consiste em atribuir a esses conteúdos uma das duas classes: positivo ou negativo. Os trabalhos realizados nesta área aplicam diferentes métodos para efectuar a detecção de polaridade em comentários de filmes e comentários de produtos, entre outros(Turney et al. [31]; Turney e Littman [32]; Pang et al. [24]). Estes trabalhos efectuam a classificação ao nível do documento.

Outra direcção na análise de sentimento consiste na identificação de objectividade/subjectividade (Hatzivassiloglou e Wiebe [14]; Wiebe et al. [34]; Wiebe [35]). Esta tarefa é commummente (Pang [23]) definida como a classificação de um determinado texto (geralmente frases) em uma das duas classes: objectivo ou subjectivo. Este problema pode por vezes ser mais difícil do que a classificação de polaridade (Mihalcea et al. [17]), a subjectividade das palavras e expressões pode depender do seu contexto e um documento objectivo pode conter frases subjectivas (por exemplo, um artigo de noticias citando opiniões de pessoas). Alem disso, como mencionado em (Su [29]), os resultados são em grande parte dependentes da definição de subjectividade utilizada quando se anotam os textos .

É neste contexto que Pais [21] propôs uma nova abordagem ao problema dos textos anotados, na maioria das vezes manualmente. Demonstrou que um corpus jornalístico e um conjunto de blogues apresentam as mesmas características de objectividade e subjectividade respectivamente, do que um corpus anotado manualmente constituído por textos objectivos e subjectivos. Como tal utilizou um corpus objectivo constituído por textos do corpus jornalístico Reuters e um conjunto de blogues como corpus subjectivo para demonstrar a similaridade dos mesmos para com um corpus constituído por textos objectivos e subjectivos (Subjectivity v1.0 corpus).

A maioria dos métodos que efectuam a análise de subjectividade e de sentimento invoca léxicos construídos manualmente ou semi-automaticamente (Yu e Hatzivassiloglou [36]; Riloff e Wiebe [25]; Kim e Hovy [16]; Turney [31]). Para a língua inglesa, estes léxicos estão disponíveis para toda a comunidade científica, mas para o português europeu não existe um recurso semelhante. A falta destes recursos condiciona bastante a aplicação de

métodos para se efectuar a análise de sentimento para a língua portuguesa, pois a disponibilidade destes léxicos permite a construção de classificadores eficientes baseados em regras de subjectividade e de sentimentos que confiam na presença do léxico de entrada no texto. Grande parte dos trabalhos realizados até à data sobre a construção de léxicos de subjectividade utilizam ferramentas avançadas de PLN como parsers sintácticos[2] (Wiebe [33]) ou ferramentas para a extracção de informação (Riloff e Wiebe [25]), ou um léxico rico e amplo como o WordNet (Esuli e Sebastiani [11]).

Contudo, os léxicos invocados por esses métodos apresentam problemas relacionados com o modo como foram implementados e com as ferramentas utilizadas.

Nesta tese é proposto construir automaticamente um léxico emocional para o português europeu, aplicando técnicas estatísticas. Como tal a aproximação seguida neste trabalho foi a de considerar a análise de sentimento direccionada para a análise de subjectividade. Os corpora em português europeu são a componente central da solução apresentada, ou seja, para não cometer o mesmo erro dos trabalhos que utilizam corpora anotados manualmente, seguiu-se os resultados obtidos por Pais [21] na demonstração da similaridade entre os corpora para a língua inglesa. Como tal, o corpus subjectivo é constituído por textos dos blogues e o corpus objectivo é constituído por textos do corpus jornalístico CETEMPúblico. Para efectuar a análise subjectiva utilizam-se os lemas por categorias morfológicas e os n-grams. Estes indícios de subjectividade são dos mais referenciados pela literatura relacionada com a análise de subjectividade, como bons indícios subjectivos. As pistas subjectivas foram extraídas dos corpora com ferramentas que efectuam essa extracção de um modo automático. Os resultados dessa extracção foram guardados em ficheiros que foram posteriormente submetidos a aplicação das técnicas estatísticas para construir o léxico de subjectividade.

1.1 Motivação e objectivos

A maioria da investigação realizada sobre a análise de sentimento até a data tem-se centrado na língua inglesa, o que é explicado pela disponibilidade de recursos para análise,

[2]Parsers sintácticos, a palavra *parsing* refere-se a *análise sintáctica*. No contexto da linguistica computacional, diz respeito a interpretação automática de frases em linguagem natural por meio de programas de computadores conhecidos como *parsers*.

tais como léxicos e corpus etiquetados manualmente. Esta investigação tem contribuído para o desenvolvimento desta área do PLN. Contudo, ainda existem partes por explorar nesta vasta área.

Assim, este trabalho pretende contribuir para o desenvolvimento desta interessante área, propondo uma nova direcção na investigação.

O objectivo deste trabalho prende-se com a necessidade de implementar uma ferramenta que possa, de forma automática, construir um léxico emocional para a língua portuguesa europeia. Para que o objectivo seja viável também se pretende mostrar que os resultados obtidos por Pais [21] na construção dos corpora, se aplicam para o português, bem como verificar o comportamento das pistas de certos trabalhos para efectuar a análise de sentimento, para o português.

1.2 Contribuições

Como já foi referido este trabalho pretende apresentar uma ferramenta que possa de forma automática construir um léxico sentimental para o português europeu.

Com o crescente fluxo de conteúdos colocados na Web nos últimos anos surgiu a necessidade de desenvolver ferramentas que possam extrair em tempo real as opiniões bem como classificar os documentos em que elas estão contidas, quanto ao seu conteúdo. É neste ideal que se enquadra este trabalho.

A análise de subjectividade utiliza vários indícios de subjectividade. A maioria dos estudos implementam metodologias distintas, mas utilizam os mesmos indícios subjectivos na maioria dos casos, sendo esses as palavras separadas em categorias morfológicas (parts-of-speech). No entanto, existe um indício de subjectividade que apresenta bons resultados. Contudo, este é pouco utilizado neste tipo de estudo. Esse indício de subjectividade é denominado por n-gram.

Assim uma das contribuições desta tese é mostrar que este tipo de indício também é bom para efectuar a análise de subjectividade.

A maior contribuição deste trabalho está relacionada com a língua para a qual se dirigiu a implementação automática deste léxico subjectivo. Isto é, as experiencias foram focalizadas para o português, no sentido de efectuar a primeira análise de sentimento para esta língua

(depois de muita pesquisa, verificou-se que não existe um recurso semelhante para a língua portuguesa europeia) podendo no entanto, aplicar este método a outras línguas.

1.3 Organização da tese

As restantes partes desta tese estão organizadas da seguinte forma: o capítulo 2, é dedicado ao enquadramento do problema proposto na área da análise de sentimento, apresentando informação sobre o que são os sentimentos e como podem ser apresentados no texto. Apresenta os problemas inerentes a construção de léxicos e apresenta a proposta de trabalho.

O capítulo 3, descreve o processo de construção dos diferentes corpora. Apresenta os indícios de subjectividade que vão ser utilizados neste trabalho, bem como as ferramentas utilizadas para extrair esses indícios.

O capítulo 4, apresenta o processo utilizado para desenvolver a construção do dicionário.

O capítulo 5, apresenta os resultados e a sua respectiva avaliação.

Finalmente, o capítulo 6 apresenta as conclusões da tese e propõe possíveis trabalhos futuros.

Capítulo 2

Trabalho Relacionado

Hoje em dia, existe cada vez mais a necessidade de classificar os conteúdos apresentados na Web, relativamente aos sentimentos neles expressos, para que qualquer individuo, empresa ou instituição possa saber em tempo real qual o sentimento inerente ao conteúdo apresentado. Esses sentimentos são em termos gerais sinónimos de subjectividade. A subjectividade pode estar expressa de diversas maneiras nos textos dos documentos. Como tal existe uma grande quantidade de metodologias distintas que a analisam.

Assim, este capítulo tem como objectivo apresentar o problema da construção de léxicos de sentimentos inerentes à análise de sentimento, na tarefa de identificação de subjectividade e ilustrar o que é mais relevante para o desenvolvimento deste trabalho. A organização é a seguinte: na secção 2.1 apresenta-se o conceito de subjectividade e sob que forma pode aparecer no texto. Na secção 2.2 são abordados diversos estudos feitos na área do PLN que utilizaram léxicos de sentimentos para efectuar a análise de sentimento. Na secção 2.3 são apresentados os problemas relacionados com as diversas implementações utilizadas para construir os léxicos. Na secção 2.4 descreve-se a proposta de trabalho para a realização deste estudo.

2.1 Subjectividade

Diariamente o ser humano é confrontado com conceitos linguísticos de um modo directo ou indirecto. Estes conceitos linguísticos surgem em qualquer tipo de relacionamento entre indivíduos, na leitura ou na escrita. Dois desses conceitos muito utilizados no dia-a-dia são a objectividade e a subjectividade.

A objectividade é a qualidade daquilo que é objectivo, externo à consciência, resultado da observação imparcial, independente das preferências individuais.

A subjectividade é o mundo interno de todo e qualquer ser humano. Este mundo interno é composto por emoções, sentimentos e pensamentos.

Analisando rapidamente estas duas definições é fácil verificar que elas representam conceitos linguísticos opostos. Quando um indivíduo utiliza a objectividade é porque pretende apresentar informação factual, isenta de qualquer tipo de emoção ou sentimento relativamente ao facto apresentado. Um bom exemplo da utilização da objectividade é as notícias dos jornais.

Quando se utiliza a subjectividade é porque se pretende em termos gerais apresentar um sentimento, emoção ou pensamento relativamente a algo. Um bom exemplo da aplicação da subjectividade é os blogues, onde os indivíduos debatem sobre produtos, dizendo se gostam ou não desses produtos.

Hoje em dia existe uma área do PLN que procura avaliar estes conceitos e classificar os textos em que eles aparecem relativamente aos sentimentos que eles apresentam. Mas antes de tentar classificar os sentimentos é necessário saber quais são os sentimentos. Em geral pode-se afirmar que esses sentimentos são entre outros, emoções, julgamentos ou ideias sugeridas por emoções. Uma emoção é constituída por um conjunto de fases: apreciação[1], alterações químicas e neurais e acções de prontidão. Uma emoção é em geral causada por uma pessoa que avalia conscientemente ou inconscientemente um evento, denominado em psicologia por apreciação. A apreciação não identifica apenas se algo é ou não positivo ou negativo, mas também denota outras medições, tais como o significado do evento, o controlo pessoal ou o envolvimento do próprio ego (Boiy [5]).

O estudo das emoções no texto pode ser realizado a partir de dois pontos de vista (Boiy [5]). Em primeiro lugar, pode-se investigar o como as emoções influenciam um escritor de textos na escolha de certas palavras e/ou outros elementos linguísticos. Em segundo lugar, pode-se investigar como o leitor interpreta a emoção num texto, e que pistas linguísticas

[1]Em inglês diz-se appraisal

são utilizadas para inferir a emoção do escritor. De seguida são apresentados os elementos linguísticos que descrevem a apreciação e as acções de prontidão que são utilizadas em textos para transmitir as emoções do autor, uma vez que eles incluem a maioria das pistas para inferir a emoção a partir do texto.

Segundo Osgood et al. [20] que investigou a forma como o significado das palavras pode ser mapeado num espaço semântico, a apreciação é constituída por três grandes dimensões: (1) avaliação positiva ou negativa, (2) o poder, controlo ou dimensão da potência e (3) uma actividade ou dimensão da intensidade. Embora estas dimensões são originalmente propostas como as dimensões dum espaço semântico, podem também ser utilizadas para organizar as categorias linguísticas da emoção ou para a detecção automática de emoções. A maior parte da investigação é dedicada à componente da apreciação das emoções.

Expressões directas, a maneira mais directa de expressar uma emoção é naturalmente expressá-la directamente. Isto pode ser feito utilizando verbos ou adjectivos, entre outros.

Elementos de acção, um excelente exemplo de acções que indicam emoções são naturalmente rir e chorar, mas podem considerar-se sinais mais subtis que demonstram emoções em determinadas circunstâncias. Um exemplo é quando se olha para o relógio quando se vê um filme, um resultado muito provável é o tédio ou falta de interesse.

Existem outras formas de expressar emoções que não se enquadram nas categorias anteriores, como a utilização de linguagem figurativa e ironia. A maioria das técnicas de classificação de sentimentos focaliza-se em termos que não demonstram realmente emoções, mas denotam avaliação, apreciação e julgamento.

2.2 Classificadores de sentimentos

Nos últimos anos tem-se verificado um crescente interesse na extracção automática ou semiautomática de opiniões, emoções e sentimentos em documentos que se encontram em suporte digital (análise de sentimento); em fornecer ferramentas e suporte para várias aplicações de PLN.

Este interesse deve-se ao crescente fluxo de conteúdos disponíveis e de fácil acesso na

Web e à necessidade de desenvolver ferramentas que sejam capazes de tratar esses mesmos conteúdos ao nível das opiniões e/ou sentimentos neles evidenciados.

A análise de sentimento automática ou semiautomática é um tópico da extracção de informação que só recentemente recebeu interesse por parte da comunidade científica. Na década de 90 do século XX, foram publicados uma série de artigos sobre o assunto. Contudo, só nos últimos anos é que se tem notado uma pequena explosão de publicações.

A ideia de efectuar esta análise de um modo automático ou semiautomático é importante para a investigação na área do marketing, onde as empresas ou instituições querem saber o que o mundo pensa sobre os seus produtos/serviços. Para acompanhar as noticias e fóruns, onde a detecção automática e rápida das opiniões é necessária, para efectuar a análise do feedback dos clientes. Esta análise também pode ser utilizada como complemento informativo para os motores de busca.

Esta forma de aplicar a análise de sentimento sobre os dados encontrados na Web é útil para qualquer empresa ou instituição, no sentido em que permite efectuar um controlo de qualidade. Antes da existência da Web, os utilizadores tinham de ser incomodados com inquéritos para dar a sua opinião sobre determinado produto ou entidade. O problema desta abordagem é que depende de muitos factores entre os quais da boa vontade do utilizador em responder ao inquérito. Este método torna-se obsoleto a partir do momento em que surge a Web porque a informação obtida nesses inquéritos pode ser obtida directamente a partir da mesma em notícias, blogues, fóruns online ou outro tipo de fonte de informação em suporte electrónico, onde por exemplo, o autor do blogue ao fazer comentários sobre determinado assunto ou ao falar sobre a sua experiencia pessoal, influência e convida os leitores a fornecer as suas próprias opiniões. Assim sendo, devido ao crescente fluxo de conteúdos colocados na Web é necessário desenvolver aplicações que extraem as opiniões em tempo real, permitindo às empresas uma reposta mais rápida às mudanças de mercado.

De um modo geral, a análise de sentimento serve para determinar a atitude de um locutor ou escritor em relação a um dado assunto. Essa atitude pode ser um juízo de valor ou avaliação, o estado emocional do autor ao escrever ou o efeito emocional que o autor pretende ter sobre o leitor.

Na análise de sentimento a tarefa básica consiste em classificar a polaridade de um determinado texto ou frase no documento, isto é, consiste em atribuir a esses conteúdos uma

das duas classes: positivo ou negativo. Os trabalhos realizados nesta área aplicam diferentes métodos para efectuar a detecção de polaridade em comentários de filmes e comentários de produtos, entre outros. Estes trabalhos efectuam a classificação ao nível do documento.

Turney et al. [31] propôs um algoritmo não supervisionado para a classificação de críticas de vários domínios na Web como recomendáveis ou não recomendáveis. Este trabalho é principalmente focalizado na utilização da orientação semântica das frases.

No inicio do processo todas as palavras são classificadas gramaticalmente com etiquetador morfossintáctico e de seguida foram identificadas todas as frases que contêm adjectivos e verbos. No próximo passo são estimadas as orientações semânticas das frases identificadas anteriormente utilizando o algoritmo PMI-IR[2]. A orientação semântica de uma frase é calculada subtraindo a sua similaridade com a palavra de referência positiva (*excelent*) a similaridade com a palavra de referência negativa (*poor*).

Turney conseguiu implementar este algoritmo recorrendo ao operador NEAR[3] do motor de busca Altavista. O algoritmo atingiu uma precisão média de 74% quando efectuou a avaliação de 410 comentários do Epinions, recolhidos a partir de diferentes domínios (comentários de automóveis, bancos, filmes e destinos de viagens). A precisão varia entre 84% para os comentários de automóveis e 66% para os filmes.

Turney e Littman [32] utilizam o mesmo algoritmo que Turney et al. [31], com a diferença de que o conjunto de palavras de referência é aumentado para 14 palavras em detrimento da duas utilizadas no trabalho de referência (sete palavras positivas e sete palavras negativas).

Pang et al. [24] aplicaram técnicas tradicionais de aprendizagem automática aos corpora de críticas de cinema tentando associar essas a classe positiva ou negativa respectivamente. Recorreram a vários algoritmos de aprendizagem automática, contudo o que mostrou ser mais eficiente foi o Support Vector Machine (SVM), atingindo uma taxa de exactidão de 82,9% com unigramas. A selecção de termos para este resultado apenas entrou em consideração com a presença de termos nos documentos em vez da frequência.

[2]PMI-IR é uma técnica estatística de PLN que utiliza os resultados da extracção de informação para calcular as relações entre palavras ou frases em relação a sua informação mútua

[3]NEAR, operador que permitia consultar o número de páginas onde duas palavras ocorressem no mesmo documento a uma distância máxima entre elas parametrizáveis.

Kim e Hovy [16] descreveram um sistema de classificação de palavras para detectar a orientação semântica de palavras opiniosas. Para tal utilizaram o WordNet [18] e três outros conjuntos de palavras positivas, negativas e neutras rotuladas manualmente.

A ideia deles é que os sinónimos das palavras positivas tendem a ser palavras positivas. Para expandir cada uma das classes de palavras escolhidas manualmente (verbos e adjectivos), utilizaram o WordNet.

Mediram a precisão do sistema utilizando uma validação cruzada 10-fold, obtendo uma precisão global (combinando positivos, negativos e neutros) de aproximadamente 77.7% para os verbos e 69.1% para os adjectivos.

Outra direcção na análise de sentimento consiste na identificação de objectividade/subjectividade. Esta tarefa é comummente (Pang [23]) definida como a classificação de um determinado texto (geralmente frases) em uma das duas classes: objectivo ou subjectivo. Este problema pode por vezes ser mais difícil do que a classificação de polaridade (Mihalcea et al. [17]), a subjectividade das palavras e expressões pode depender do seu contexto e um documento objectivo pode conter frases subjectivas (por exemplo, um artigo de notícias citando opiniões de pessoas). Alem disso, como mencionado em (Su [29]), os resultados são em grande parte dependentes da definição de subjectividade utilizada quando se anotam os textos.

Hatzivassiloglou e Wiebe [14] propuseram um método que combina estatisticamente dois indicadores de graduability para estudar o comportamento de adjectivos dinâmicos, adjectivos com orientação semântica e o grau do adjectivo, num classificador simples de subjectividade.

O processo começa com a extracção de todos os adjectivos com frequência igual ou superior a 300 do corpus. De seguida são atribuídas automaticamente etiquetas de graduability a cada palavra. Das 496 palavras resultantes, apenas 453 vão ser consideradas.
Utilizando uma validação cruzada 4-fold treinada com três quartos dos 453 adjectivos, utilizando os restantes para teste em cada fold, obtiveram uma precisão de 88.08% para o conjunto de adjectivos definidos.

Wiebe et al. [34] desenvolveu um sistema automático para efectuar a etiquetagem subjectiva. Aplicou uma validação cruzada 10-fold ao corpus. Em cada fold, um conjunto é

utilizado para teste e os outros nove são utilizados para treino. A selecção de características, o modelo de selecção, bem como a estimativa dos parâmetros são realizadas novamente em cada fold. É incluída uma característica binária para cada um dos seguintes casos: a presença na frase de um pronome, um adjectivo, um número, um modal diferente de *will* (modo verbal inglês), e um advérbio diferente de *not*. Também incluíram uma característica binária representando ou não a frase que começa um novo parágrafo. E por último, é incluído uma característica que representa a co-ocorrência de tokens de palavras e sinais de pontuação com a classificação objectiva e subjectiva.

A aplicação de um classificador probabilístico obteve uma precisão media de 72.17% na etiquetagem subjectiva, 20 pontos percentuais acima da precisão de referência obtido no facto de escolher sempre a classe mais frequente.

Wiebe [35] propôs um método simples para identificar automaticamente indícios de subjectividade em agrupamentos de palavras (collocations) no texto.

No princípio o método é utilizado para identificar agrupamentos de palavras compostas por um número fixo de palavras que, quando aparecem juntas no texto, tendem a ser subjectivas. Em seguida, o método é novamente aplicado para os mesmos dados, mas com todas as palavras que aparecem só uma vez substituído pela palavra *UNIQUE*. As duas aplicações obtiveram resultados diferentes, no entanto, ambos mostram um aumento de precisão.

Pang e Lee [22] implementaram um detector de frases subjectivas. A figura 2.1 ilustra o detector de subjectividade. Estudos prévios demonstraram que o conjunto de frases de conteúdo subjectivo apresenta a informação sobre o sentimento do texto de uma forma mais compacta.

O conjunto de treino do detector de subjectividade é constituído por 5000 pedaços de críticas tais como *ousado, imaginativo* do site *www.rottentomatoes.com*. E por 5000 frases de resumos de filmes do site *www.imdb.com*. A tarefa de aprendizagem do detector foi realizada com o método do corte mínimo em grafos. Com este método partiram do pressuposto que frases próximas umas das outras tendem a pertencer à mesma classe (objectiva ou subjectiva). Utilizaram o algoritmo Naive Bayes como classificador de polaridade, ficando demonstrado que a classificação foi mais eficiente com as frases subjectivas extraídas do que com a utilização dos textos completos. Utilizando um SVM atingiram uma taxa de exactidão de 87,2%.

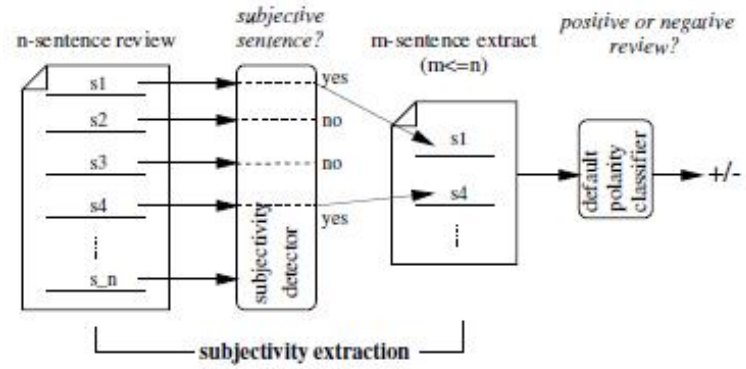

Figura 2.1: Categorização da polaridade recorrendo à detecção de subjectividade.

2.3 Criação de dicionários de sentimentos

Um dicionário (ou léxico) é uma lista de palavras ou expressões com meta dados[4] arbitrários anexados. Este difere dos dicionários tradicionais, na medida em que é desenvolvido para a leitura efectuada pelos computadores ao contrário da leitura feita por humanos. Além disso, não é necessário colocar a informação das etiquetas gramaticais[5]. No estudo da análise de sentimentos muitos investigadores optam por criar léxicos para fornecer diferentes elementos onde são capazes de detectar a presença de opiniões ou factos.

A maioria dos métodos que efectuam a análise de subjectividade e de sentimento invoca léxicos construídos manualmente ou semi-automaticamente (Yu e Hatzivassiloglou [36]; Riloff e Wiebe [25]; Kim e Hovy [16]; Turney [31]). Para a língua inglesa, estes léxicos estão disponíveis para toda a comunidade científica, mas para o português europeu não existe um recurso semelhante. A falta destes recursos condiciona bastante a aplicação de métodos para se efectuar a análise de sentimento para a língua portuguesa, pois a disponibilidade destes léxicos permite a construção de classificadores eficientes baseados em regras de subjectividade e de sentimentos que confiam na presença do léxico de entrada no texto. Grande parte dos trabalhos realizados até à data sobre a construção de léxicos de subjectividade utiliza ferramentas avançadas de PLN como parsers sintácticos[6] (Wiebe [33]) ou

[4]Os meta dados são dados sobre outros dados.

[5]As etiquetas gramaticais são utilizadas para assinalar no texto as categorias gramaticais das palavras (verbo, adjectivo, entre outros.), em inglês diz-se *part-of-speech tag*.

[6]Parsers sintácticos, a palavra *parsing* refere-se a *análise sintáctica*. No contexto da linguistica computacional, diz respeito a interpretação automática de frases em linguagem natural por meio de programas

ferramentas para a extracção de informação (Riloff e Wiebe [25]), ou um léxico rico e amplo como o WordNet (Esuli e Sebastiani [11]).

Hatzivassiloglou e McKeown [13] distinguiram os adjectivos positivos dos negativos através da hipótese que os adjectivos agrupados pela conjunção *e* tendem a ser similares, e tendem a ser dissimilares se forem agrupados pela conjunção *mas*.

Para realizar a validação da hipótese utilizaram um léxico de adjectivos com orientação semântica[7] pré-estabelecida. Depois de ter demonstrado que a hipótese era válida, e que as conjunções entre os adjectivos fornecem indirectamente a informação sobre a orientação semântica dos adjectivos, procuraram grupos de adjectivos com a mesma orientação.

Turney [31], construiu um léxico de adjectivos polares. O seu algoritmo começa com um pequeno conjunto de palavras sementes[8] relativamente as quais se conhece a polaridade (*2 palavras: excellent e poor*). Depois, dado um conjunto de termos com orientação semântica desconhecida, ele utiliza o algoritmo PMI-IR[9] (Turney [30]) para formular queries para a Web e determinar para cada termo, a sua Pointwise Mutual Information (PMI)[10] com as duas sementes, isto para todo o vasto conjunto de documentos presentes no motor de busca AltaVista. A orientação semântica de um termo é então determinada a partir da diferença entre a sua associação PMI com a semente positiva *excelent* e a sua associação PMI com a semente negativa *poor*. A lista resultante de termos e respectiva orientação sentimental foi utilizada para implementar um classificador de polaridade ao nível do documento. Utilizando um algoritmo não supervisionado obteve uma taxa de exactidão de 66% na classificação de críticas de cinema.

Turney e Littman [32] começam por definir um conjunto de adjectivos objectivos e subjectivos. Utilizam o mesmo algoritmo que Turney et al. [22], com a diferença que o

---

de computadores conhecidos como *parsers*.

[7]Orientação semântica, refere-se a polaridade das palavras ou das expressões. Essa polaridade pode ser positiva, negativa ou neutra.

[8]Palavras sementes, são um conjunto de palavras que servem de base para criar um novo conjunto de palavras.

[9]PMI-IR é uma técnica estatística de PLN que utiliza os resultados da extracção de informação para calcular as relações entre palavras ou frases em relação a sua informação mútua

[10]é uma medida da área de Teoria da Informação, utilizada para medir a relação entre uma ou mais palavras, dentro de um conjunto de textos.

conjunto de palavras de referência é aumentado para 14 palavras em detrimento das duas utilizadas no trabalho de referência (sete palavras positivas e sete palavras negativas). Ao fazê-lo, eles também determinam a orientação semântica de uma palavra baseando-se na PMI e Latent Semantic Analysis (LSA)[11] (Turney e Littman [32]).

Yu e Hatzivassiloglou [36] apresentaram um sistema de detecção e classificação de opiniões como positivas, negativas ou neutras. O sistema é utilizado para classificar documentos bem como frases.

Para determinar se a opinião presente numa frase é positiva, negativa ou neutra eles utilizaram o léxico de palavras sementes similares produzido por Hatzivassiloglou e McKeown [13] e é expandido para construir um conjunto maior de palavras com orientação semântica similar com um método semelhante ao método proposto por Turney [31].

A detecção e classificação de opiniões ao nível da frase alcançou uma precisão de 91%.

Kim e Hovy [16] descreveram um sistema de classificação de palavras para detectar a orientação semântica de palavras opiniosas. Para tal utilizaram o WordNet [18] e três outros conjuntos de palavras positivas, negativas e neutras rotuladas manualmente.

A ideia deles é que os sinónimos das palavras positivas tendem a ser palavras positivas. Para expandir cada uma das classes de palavras escolhidas manualmente (verbos e adjectivos), utilizaram o WordNet.

Mediram a precisão do sistema utilizando uma validação cruzada 10-fold, obtendo uma precisão global (combinando positivos, negativos e neutros) de aproximadamente 77.7% para os verbos e 69.1% para os adjectivos.

Esuli e Sebastini [10] propuseram um método de aprendizagem semi-supervisionado para determinar a orientação semântica dos termos. A novidade do método reside no facto deles explorarem uma fonte de informação que técnicas anteriores nunca tentaram usar para resolver esta tarefa, ou seja, as definições textuais que os termos têm num dicionário online.

O seu pressuposto básico é que os termos com orientação semelhante tendem a ter uma definição semelhante, por exemplo: a definição de *honest* e *intrepid* irão ambas ter a expressão *appreciative*, enquanto a definição de *disturbing* e *superfluous* irão ter ambas a expressão *derogative*.

[11] Latent Semantic Analysis é uma técnica estatística de PLN que análise as relações entre um conjunto de documentos e os termos que contêm.

O conjunto de treino começa com um pequeno conjunto de sementes positivas e negativas que é enriquecido através da pesquisa efectuada num dicionário. Os termos positivos do conjunto são acrescentados a partir dos termos relacionados com eles através das relações que indicam uma orientação semelhante, os sinónimos, e das relações que indicam uma orientação oposta, os antónimos. Os termos negativos do conjunto de treino são acrescentados através de um processo análogo.

Mihalcea et al. [17], descrevem experimentos para a classificação de subjectividade para o romeno. Começam com um léxico inglês subjectivo com 6856 palavras, do Opinion-Finder[12] (Wiebe e Riloff [25]), e traduzem-nos automaticamente para o romeno utilizando dois dicionários bilingues, obtendo um léxico romeno com 4983 palavras.

A avaliação manual de uma amostra de 123 palavras deste léxico mostrou que 50% das palavras indiciam subjectividade.

Banea et al. [1], propuseram um método para a criação de um léxico subjectivo para línguas com escassos recursos, utilizando um pequeno conjunto de palavras sementes subjectivas, um dicionário online e um pequeno corpus em cru[13], juntamente com um processo de bootstrapping[14] que classifica novas palavras candidatas com base numa medida de similaridade. O processo começa com um pequeno conjunto de sementes de palavras subjectivas obtidas manualmente, e com a ajuda de um dicionário online, produz um léxico subjectivo de potenciais candidatos. O léxico obtido após cinco iterações do método utilizado para a classificação de sentimento ao nível da frase, indica uma melhoria de 18% relativamente ao léxico de Mihalcea et al. [17]. Os candidatos são então classificados com base na medida de semelhança LSA e as primeiras 4000 palavras foram utilizadas para construir um classificador de subjectividade ao nível da frase. Utilizando um classificador não supervisionado ao nível da frase, alcançaram uma medida F-mesure[15] de 62% ao nível da subjectividade.

[12]Opinion-Finder, é um sistema que efectua automaticamente a análise subjectiva para o inglês. Identifica quando é que as opiniões, sentimentos, especulações e estados privados estão presentes no texto.

[13]Corpus em cru, é um corpus que ainda não foi processado.

[14]Bootstrapping, é um método que expande um pequeno conjunto de sementes num grande conjunto de palavras ou expressões. Em cada iteração o conjunto de sementes é completado com palavras relacionadas encontradas num dicionário online, e são filtradas por uma medida de similaridade.

[15]Medida de performance utilizada na extracção de informação, é definida como uma medida harmónica entre a precisão e a cobertura (Recall)

Jikoun e Hofmann [15], propuseram um método para criar um léxico subjectivo para o holandês com base num léxico inglês, um tradutor online e o léxico WordNet [11].

O método começa com um léxico inglês de palavras positivas e negativas, traduzidas automaticamente para a língua alvo (neste caso para o holandês). Depois aplicam um algoritmo de PageRank[16] ao WordNet holandês para filtrar e expandir o conjunto de palavras obtida através da tradução. O algoritmo de PageRank vê o léxico WordNet como um grafo onde as palavras e os conceitos estão conectados por relações tais como *sinónimos, antónimos, hipônimos, entre outros*.

Os melhores resultados foram obtidos quando o método utilizou apenas as relações: *sinónimos* e *antónimos*, na classificação simultânea de palavras positivas e negativas. O método alcança uma precisão de 82% nas 3000 primeiras palavras negativas e 62% nas 3000 primeiras palavras positivas.

2.4 Problemas

Os léxicos de sentimentos construídos até a data para efectuar o PLN têm vários problemas associados, apesar de terem permitido obter bons resultados nos estudos científicos em que foram aplicados. Esses problemas estão directamente relacionados com o modo como eles foram implementados, isto é, as ferramentas utilizadas bem como os conhecimentos linguísticos associados.

Na construção de um léxico de sentimento não se deve utilizar a informação associada ao conhecimento linguístico, pois a utilização dessa informação influencia directamente os resultados, visto que são dirigidos pela língua. Quando se fala em conhecimentos linguísticos, fala-se em tudo o que esta associado a polaridade, subjectividade, dicionários online, entre outros.

A maioria dos trabalhos que direccionam o seu estudo para outra língua que não o inglês

[16]PageRank, é uma família de algoritmos de análise de rede que dá pesos numéricos a cada elemento de uma colecção de documentos hiperligados, como por exemplo as páginas da Internet, com o propósito de medir a sua importância nesse grupo por meio de um motor de busca

tendem a utilizar dicionários subjectivos implementados para a língua inglesa como fonte de conhecimento para construir o léxico a ser utilizado nas suas respectivas investigações (Mihalcea et al. [17], Banea et al. [1], Jikoun e Hofmann [15]). Esses dicionários já demonstram um bom desempenho na classificação de subjectividade ou polaridade, e como são bons nessas tarefas, são traduzidos para a língua alvo do estudo para servir de base de conhecimento.

A utilização dos dicionários online influencia directamente os resultados da construção dos léxicos, pois um dicionário fornece todo o tipo de informação sobre qualquer palavra. Basta ver a definição da palavra dicionário para compreender que a sua utilização influencia o comportamento da construção, pois estes fornecem informação sobre as categorias gramaticais (adjectivos, verbos, entre outros) e outros aspectos gramaticais (género, número, entre outros), entre outras informações.

Esuli e Sebastini [11] utilizam um dicionário online e as definições textuais das palavras para expandir um conjunto de treino manualmente construído de palavras positivas e negativas.

Chesley et al. [6] utilizam a informação da classe dos verbos e um dicionário online para classificar o sentimento expresso nos adjectivos e nos blogues.

Outro problema que surge na construção de léxicos de sentimentos está associada com aqueles trabalhos que pegam num conjunto de palavras semente para expandir o léxico. Isto é, por exemplo, nos trabalhos implementados por Turney [31] e Turney e Litman [32], eles utilizam um pequeno conjunto de palavras sementes, relativamente às quais se tem conhecimento da sua polaridade para construir o léxico.

Estes tipos de trabalhos utilizam métodos de similaridade para expandir os conjuntos de palavras sementes iniciais, e consequentemente as palavras que vão ser escolhidas vão estar directamente relacionadas com o facto de saber que as palavras semente tinham uma dada polaridade.

Hatzivassiloglou e McKeown [13] utilizam heurísticas para extrair adjectivos dos textos para classificar o sentimento expresso em textos.

Em muitos estudos, são utilizados juízos humanos com formação no estudo da linguagem natural para construir manualmente o dicionário de sentimentos. Essa forma de construir os léxicos é uma ferramenta que influencia consideravelmente os resultados, pois estes juízes indicam as palavras que apresentam na perfeição a informação necessária para o léxico que se pretende construir para efectuar o estudo.

Quando se pretende construir um dicionário para o PLN existem técnicas tão eficientes como as que foram referidas aqui, só que tem uma grande vantagem relativamente a essas, é que não influencia os resultados com dados que não fazem parte dos textos dos corpora.

2.5 Proposta de trabalho

Pais [21] utilizou dois métodos diferentes para avaliar a hipótese de que os textos da Wikipédia são objectivos e os textos dos blogues são subjectivos. Os métodos utilizados foram o Método de Rocchio e o Modelo da Linguagem. O objectivo desses métodos foi a de avaliar a similaridade entre o corpus constituído por textos objectivos e subjectivos (Subjectivity v1.0 corpus [22]) anotado manualmente ao nível morfológico e a Wikipédia e os blogues. Também se aplicou o mesmo processo ao corpus jornalístico Reuters com o intuito de verificar a que classe pertence.

Na avaliação utilizou a informação das categorias morfológicas para ver qual o seu comportamento em termos de similaridade entre os corpora. Ambos os métodos utilizaram uma validação cruzada *10-fold*, onde se construíram dez conjuntos de treino, para cada corpus em estudo, contendo 90% das frases da Wikipédia, dos blogues e do Reuters respectivamente; e um conjunto de teste, para cada classe (subjectiva e objectiva) contendo 10% das frases subjectivas e objectivas do corpus Subjectivity v1.0.

No método de Rocchio utilizou a medida de *similaridade Cosine* para calcular o ângulo entre dois vectores, os vectores de treino e de teste.

A tabela 2.1 mostra a *similaridade de Cosine* para cada categoria morfológica na avaliação da similaridade entre a Wikipédia e o corpus Subjectivity v1.0.

Morphological Level	Subjective	Objective	Class
All Words	0.76	0.79	Objective
All ADJ	0.54	0.61	Objective
All V	0.71	0.67	Subjective
All N	0.66	0.69	Objective
All ADJ + All V	0.65	0.66	Objective
All ADJ + All N	0.65	0.68	Objective
All N + All V	0.70	0.69	Subjective
All ADJ + All N + All V	0.68	0.69	Objective

Tabela 2.1: Resultados da Wipédia[21].

Verifica-se que algumas categorias tem comportamento subjectivo, mas quando se olha para o conjunto como um todo, todas as palavras (*all words*), o comportamento da Wikipédia é objectivo.

A tabela 2.2 mostra a *similaridade de Cosine* para cada categoria morfológica na avaliação da similaridade entre os blogues e o corpus Subjectivity v1.0.

Morphological Level	Subjective	Objective	Class
All Words	0.60	0.56	Subjective
All ADJ	0.52	0.49	Subjective
All V	0.53	0.48	Subjective
All N	0.47	0.43	Subjective
All ADJ + All V	0.49	0.48	Subjective
All ADJ + All N	0.48	0.44	Subjective
All N + All V	0.50	0.45	Subjective
All ADJ + All N + All V	0.47	0.46	Subjective

Tabela 2.2: Resultados dos blogues[21].

Verifica-se que todas as categorias morfológicas tem comportamento subjectivo e consequentemente o corpus dos blogues é subjectivo.

A tabela 2.3 mostra a *similaridade de Cosine* para cada categoria morfológica na avaliação da similaridade entre o Reuters e o corpus Subjectivity v1.0.

Morphological Level	**Subjective**	**Objective**	**Class**
All Words	0.64	0.68	Objective
All ADJ	0.30	0.40	Objective
All V	0.38	0.37	Subjective
All N	0.34	0.47	Objective
All ADJ + All V	0.36	0.38	Objective
All ADJ + All N	0.35	0.49	Objective
All N + All V	0.36	0.47	Objective
All ADJ + All N + All V	0.37	0.47	Objective

Tabela 2.3: Resultados do Reuters[21].

Verifica-se que uma das categorias tem comportamento subjectivo, mas quando se olha para o conjunto como um todo, todas as palavras (*all words*), o comportamento do Reuters é objectivo.

Para aplicar o método do Modelo da linguagem utilizou a ferramenta *CMU-ToolKit*[17]. Assim, nesse contexto, o menor valor de *perplexity (Px)* e de *entropia (H)* indica qual a classe a que pertence o conjunto.

		Model		
		Wikipedia	Weblogs	Reuters
Text	Objective	Px = 691.27 H = 9.43	Px = 2027.06 H = 10.99	Px = 1104.03 H = 10.11
	Subjective	Px = 880.67 H = 9.75	Px = 1991.09 H = 10.96	Px = 1226.34 H = 10.26

Tabela 2.4: Resultados obtidos pelo Modelo da linguagem[21].

Constata-se na tabela 2.4 que no caso da Wikipédia e do Reuters os valores apresentados

[17]CMU-ToolKit, é um conjunto de software unix concebido para facilitar a investigação efectuada pela comunidade científica quando trabalha com o Modelo da Linguagem.

são menores para a classe objectiva, ao passo que no caso dos blogues os menores valores apontam para a classe subjectiva.

Os métodos obtiveram resultados diferentes, contudo em ambos os casos verificou-se que a hipótese era verdadeira e assim os textos da Wikipédia representam um corpus objectivo bem como os textos dos blogues representam um corpus subjectivo, bem como o corpus jornalístico Reuters representa um corpus objectivo.

Estes resultados permitiram demonstrar que estes tipos de corpora tem vantagens para a comunidade científica, pois eles podem ser obtidos de forma automática, não são anotados manualmente, podem ser construídos em qualquer altura, para qualquer língua e são bastantes abrangentes.

A proposta de trabalho consiste em pegar nos resultados obtidos por Pais [21] na identificação da diferença entre a subjectividade e a objectividade com os blogues e a Wikipédia, e o facto de ter demonstrado que um corpus jornalístico pertence à classe objectiva, demonstrou isso com corpus jornalístico Reuters, para construir um léxico emocional.

Os resultados do estudo científico apresentado nesta secção mostram que os blogues podem ser utilizados como corpus subjectivo e a Wikipédia e o corpus jornalístico Reuters podem ser utilizados como corpus objectivo para a língua inglesa.

Assim, com base nesses resultados propomos construir automaticamente um léxico emocional para a língua portuguesa europeia, aplicando técnicas estatísticas e utilizando como corpora um corpus subjectivo constituído por textos de blogues portugueses e um corpus objectivo constituído por textos do corpus jornalístico CETEMPúblico.

A construção automática de um léxico emocional a partir de corpora que se sabem à partida serem subjectivos/objectivos ou positivos/negativos, é uma forma de contornar grande parte dos problemas associados à utilização de conhecimentos linguísticos e relativamente às ferramentas utilizadas. Pois, apesar de se saber qual a orientação dos corpora, não existe nada nos textos dos mesmos que diga directamente que este ou aquele conjunto de palavras são úteis para a construção do léxico, isto é, não há etiquetas a dizer que uma dada palavra é subjectiva, objectiva, positiva ou negativa. E assim os léxicos obtidos apenas apresentam

o facto de este ou aquele conjunto de palavras poder vir a ser subjectivo/objectivo ou positivo/negativo respectivamente.

O automatismo do processo remove a ambiguidade associada ao julgamento dos termos por parte de avaliadores humanos com conhecimentos linguísticos em linguagem natural.

Com este mecanismo evita-se que a construção do léxico de sentimentos seja dirigida pela língua.

Capítulo 3

Corpora

Um corpus (plural: corpora) é um conjunto de dados linguísticos reais criteriosamente colectados, utilizados em diversas áreas. Deve ser constituído por dados autênticos, legíveis por computador e representativos de uma língua ou variedade de línguas da qual se pretende efectuar o estudo.

Na maioria das aplicações PLN os corpora são a base de funcionamento do processo. No estudo da análise de sentimentos a utilização de corpora é fundamental. Normalmente, os corpora utilizados são anotados manualmente por vários especialistas no domínio a que os documentos se referem. Contudo, como mencionado em Pais [21], estes corpora anotados manualmente são uma desvantagem, porque em primeiro lugar estes são limitativos, uma vez que são pequenos e por isso abrangem um número restrito de assuntos. Em segundo, ao serem anotados manualmente estes corpora apresentam indirectamente conhecimentos linguísticos associados à avaliação efectuada pelos juízes humanos.

Como já foi referido nesta tese, pretende-se efectuar uma análise de subjectividade que permite a implementação de um léxico subjectivo. Como tal são necessários corpora subjectivos e objectivos. Os problemas dos corpora já existentes, apresentados anteriormente, obrigam a procurar de uma melhor forma de obtenção dos corpora.

Pais [21] ultrapassou este problema mostrando que um corpus jornalístico pode ser utilizado como corpus objectivo e que um conjunto de blogues pode ser utilizados como corpus subjectivo. Face aos resultados obtidos por Pais [21] para o inglês, a base para construção dos corpora em português para a realização deste trabalho vai ser a mesma.

Assim vão ser utilizados dois tipos diferentes de corpora em português europeu, um corpus com textos objectivos (origem: um corpus jornalístico, CETEMPúblico) e um corpus com textos subjectivos (origem: um conjunto de blogues).

Para efectuar a análise de subjectividade são necessários bons indícios de subjectividade. Depois de muita pesquisa e leitura sobre esta área, verificou-se que os indícios de subjectividade que apresentavam melhor resultados na identificação de subjectividade foram as categorias morfológicas[1] bem como os n-grams[2]. Ora também se constatou que os corpora utilizados neste tipo de análise também eram anotados manualmente ao nível morfológico, no caso das categorias morfológicas.

Como tal para contornar este problema utilizou-se uma ferramenta para efectuar automaticamente a etiquetagem ao nível morfológico dos corpora. Também se utilizou uma ferramenta que extrai automaticamente os n-grams.

Neste capítulo vai ser apresentada a história dos corpora linguísticos . Mas antes, é necessário definir o que é um corpus. Também vão ser apresentados os indícios de subjectividade utilizados para efectuar a análise de subjectividade que permite efectuar a construção do léxico. Bem como as ferramentas utilizadas para extrair esses indícios. A organização é a seguinte: na secção 3.1 defina-se o que é um corpus. Na secção 3.2 apresenta-se a história dos corpora. Na secção 3.3 apresenta-se a informação sobre a extracção dos corpora. As secções 3.3.1 e 3.3.2 apresentam os corpora utilizados neste estudo. A secção 3.4 apresenta o pré-processamento que foi necessário aplicar aos corpora antes de aplicar qualquer técnica de extracção de unidades lexicais. Também são apresentadas as unidades lexicais utilizadas para efectuar a construção do léxico, bem como as ferramentas utilizadas para extrai-las, 3.4.1, 3.4.2.

3.1 O que é um corpus?

Um corpus [19] [7] é simplesmente descrito como um grande corpo de evidências linguísticas tipicamente compostos por usos atestados da linguagem. Podendo-se contrastar esta forma de evidência linguística com frases que não foram criadas como resultado da comunicação no contexto, mas sim sobre a reflexão meta linguística mediante o uso linguístico,

[1] categorias gramaticais

[2] Sequência de palavras contíguas

o tipo de dados comum na aproximação geradora da linguística. Os dados do corpus não são compostos por reflexões teóricas. É composto por uma variedade de material de conversas diárias (*Ex: a secção da fala do Britsh National Corpus*[3]), notícias da rádio (*Ex: IBM/Lancaster Spoken English Corpus*[4]), publicações escritas (*Ex: a maioria da secção escrita do Britsh National Corpus*[5]) e escritos de jovens crianças (*Ex: o Leverhulme Corpus of Children´s Writing*[6]). Tais dados são recolhidos em conjunto como os corpora que podem ser utilizados para uma grande proposta de investigação. Tipicamente estes corpora são legíveis por computadores tentando explorar os recursos baseados no papel da linguística ou gravações de áudio que correm milhões de palavras, o que é impraticável. Assim, enquanto os corpora podem basear-se no papel, ou mesmo simplesmente nas gravações de áudio, a ideia tida aqui é que os corpora são legíveis por computadores. Os corpora têm usos tanto na linguística como no PLN, e são de interesse para os investigadores de áreas, tal como a estatística literária. Os corpora são um recurso multifuncional.

Com este começo, necessita-se uma definição de corpus simplesmente mais refinada do que aquela que foi introduzida até ao momento. Foi estabelecido que um corpus é colecção de dados que ocorrem naturalmente na linguagem. Mas qualquer colecção de dados da linguagem que contenha entre três frases e três milhões de palavras pode ser considerado um corpus? O termo corpus só deve ser aplicado adequadamente a uma colecção de dados bem organizada, colecções dentro dos limites de uma amostragem significativa concebida para explorar uma determinada característica linguística (ou um conjunto de características) através da colecção de dados. A amostragem é de uma importância crucial na concepção do corpus. A amostragem é inevitável. Salvo no caso em que o objecto de estudo seja altamente restrito às sub-linguas ou línguas mortas, é absolutamente impossível recolher todas as emoções da linguagem natural juntamente com um corpus. Como consequência, o corpus deve ter como objectivo o equilíbrio e a representatividade dentro da amostragem específica, a fim de permitir uma linguagem variada a ser estudada ou modelada. A melhor maneira para explicar esses termos é através de um exemplo. Imagine que um investigador tem a tarefa de desenvolver um gestor de diálogos para um sistema de venda de bilhetes por telefone e decide construir um corpus para ajudar nesta tarefa. A amostragem aqui é evidentemente a relevância dos dados associados à venda de bilhetes por telefone para a

[3]http://www.natcorp.ox.ac.uk/

[4]http://icame.uib.no/lanspeks.html

[5]http://www.natcorp.ox.ac.uk/

[6]http://www.lancs.ac.uk/fass/projects/lever/index.htm

construção do corpus. Seria totalmente inadequado para amostragem utilizar os romances da Jane Austen ou as conversas espontâneas cara à cara, para o efeito de assumir a tarefa de modelar os diálogos baseados nas transacções efectuadas ao telefone. No domínio da venda dos bilhetes por telefone, pode haver uma variedade de bilhetes diferentes vendidos, exigindo cada uma a colocação de diferentes perguntas. Assim, pode argumentar-se que existem várias categorias linguísticas distintas de venda de bilhetes. Assim, o corpus é equilibrado pela inclusão de uma vasta quantidade de tipos de conversas telefónicas de vendas de bilhetes nele, com os tipos organizados em subpartes coerentes (*Ex: venda de bilhetes de comboio, venda de bilhetes de avião*). Finalmente, dentro de cada uma destas categorias pode haver um pequeno ponto na gravação da conversa, ou mesmo apenas as conversas dos operadores de venda. Se fosse efectuado a gravação de apenas uma conversa poderia ser altamente exclusivo. Se fosse apenas uma gravação das conversas atendidas por um operador, não se poderia ter a certeza de que elas seriam idênticas para todos os operadores. Consequentemente, o corpus tem como objectivo a representatividade incluindo nele um conjunto de oradores, a fim de que as individualidades sejam excluídas.

3.2 História de um corpus linguístico

É difícil delinear uma história do corpus linguístico. Nesta forma moderna informatizada, são só conhecidos desde o final da década de 1940. A ideia básica de utilizar usos atestados da linguagem para o estudo da linguagem era claramente prévio a esta data, mas o problema era que a recolha de grandes volumes de dados linguísticos na época pré-computador era tão difícil como quase impossível. Houve exemplos notáveis que foram alcançados através da utilização de uma grande quantidade de mão-de-obra (Kaeding é um exemplo notável disso). No entanto, na realidade, corpus lingusiticos na forma que os conhecemos hoje, onde qualquer utilizador de PC, pode com relativa facilidade, explorar os corpora com milhões de palavras, é um fenómeno muito recente.

A ligação crucial entre computadores e a manipulação de grandes quantidades de evidências linguísticas foi esquecido por Bussa no fim dos anos 40. Durante a década de 1950 o primeiro grande projecto para a construção de corpora comparáveis foi realizado por Juilland, que também articulou claramente os conceitos por detrás das ideias de amostragem, o equilíbrio e a representatividade. A implantação de corpus de corpus linguístico em inglês

começa no final dos anos 50, com um trabalho nos Estados Unidos no Brown Corpus[7] e o trabalho na Grã-Bretanha sobre o levantamento de usos ingleses. Trabalhos nos corpora linguísticos ingleses cresceram em particular em toda a década de 1960, 1970 e 1980, com importantes marcos como um corpus transcrito da linguagem falada, um corpus com etiquetagem morfossintáctica manual, um corpus com etiquetagem morfossintáctica automática. Durante a década de 1980, o número de corpora disponíveis cresceu continuamente bem como o tamanho desses corpora. Esta tendência tornou-se clara na década de 1990, com corpora como o British National Corpus e o Bank of English atingindo uma grande dimensão (100 milhões de palavras e 300 milhões de palavras do inglês moderno), que teria sido a todos os efeitos práticos, impossível na época pré-computadores. A outra tendência que se tornou perceptível durante a década de 1990 foi a natureza crescente de corpora linguísticos multilíngua.

3.3 Extracção dos Blogues e do CETEMPúblico

Sabendo que o CETEMPúblico e os blogues podem ser utilizados como corpus objectivo e subjectivo respectivamente, já se pode construir os corpora que vão ser utilizados neste trabalho.

Quando se pretende construir um corpus aparece um problema associado a essa tarefa, que consiste em saber qual é a melhor dimensionalidade do corpus, para ter um corpus que seja estatisticamente significante.

O problema da dimensionalidade do corpus surge quase sempre quando se pretende analisar este tipo de conteúdo, pois no caso de se pretender utilizar amostras muito grandes, estas demorariam muito tempo a processar e não permitiriam retirar conclusões em tempo útil. Por outro lado, não se pode utilizar amostras muito pequenas, pois podem induzir a erro.

Para contornar este problema de dimensionalidade, Pais [21] determinou o erro ao testar várias dimensões de amostras, de modo a saber qual a melhor proporção de dados a utilizar para construir os respectivos corpora. Utilizou a equação 3.1 para calcular o erro [2].

[7] http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

$$n = p(1-p)\left[\frac{Z_{\alpha/2}}{d}\right]^2 \tag{3.1}$$

onde: d - Representa o diâmetro máximo do erro permitido; p - Representa a probabilidade de se observar um evento a partir da distribuição de probabilidade (p = 0.5); $Z_(\alpha/2)$ - Representa o inverso da distribuição norma; α - Representa o nível de significância, que é adicional ao nível de confiança (α = 0.001); n - Representa o tamanho da amostra

Obteve um erro aceitável de 0.0003592637 em relação a distribuição normal $\mathcal{N}(0,1)$, para uma dimensão de tamanho igual a 100 Megabytes (MB) de texto.

Assim sendo, vão ser utilizados 100 MB extraídos aleatoriamente de cada corpus para construir os corpora (CETEMPúblico/ Blogues).

3.3.1 Corpus CETEMPúblico

O CETEMPúblico é um corpus de aproximadamente 180 milhões de palavras em português europeu, constituído por textos extraídos de aproximadamente 2600 edições do jornal Público entre os anos de 1991 e 1998 na versão 1.0.

Foi criado pelo projecto processamento computacional do português (projecto que deu origem à Linguateca) após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal PÚBLICO em Abril de 2000.

Depois de efectuar o download da versão 1.7, seleccionamos aleatoriamente 100 MB de textos divididos por vários ficheiros.

3.3.2 Corpus blogues

Como não existe nenhum repositório de blogues para fazer download dos mesmos, solucionou-se o problema, implementando um programa ao qual se indicou os domínios para fazer o respectivo download.

Os diferentes domínios abrangem uma série diferente de conteúdos que vão do desporto, à politica passando por ambientalismo e religião, entre outros temas, num total de 19 domínios diferentes.

Como já foi referido anteriormente só foram seleccionados aleatoriamente 100 MB de textos divididos por vários ficheiros.

Na tabela 3.1 é apresentada mais informação sobre os 100 MB de textos extraídos aleatoriamente dos corpora.

Corpora	**CETEMPúblico**	**Blogues**
Frases	274804	1240872
Palavras	6191068	16025818

Tabela 3.1: Dimensão do corpora.

3.4 Computação dos Corpora

Depois de apresentar na secção 3.3 a informação sobre os corpora, é necessário numa fase prévia à aplicação da metodologia, efectuar o processamento dos corpora para que possam ser utilizados de forma eficiente pelo método.

Antes da aplicação de qualquer técnica de extracção das unidades lexicais, foi necessário efectuar uma formatação dos corpora bem como uma limpeza dos mesmos. Esta tarefa incidiu principalmente no corpus dos blogues. A limpeza serviu para remover todos os rótulos de formatação (código HTML e XML); a formatação consiste na delimitação das frases, parágrafos e títulos por uma terminologia semelhante ao código *XML*. As frases aparecem delimitadas por $< S >$, inicio da frase e $< /S >$, fim da frase; os parágrafos e os títulos têm os mesmos delimitadores com a diferença de aparecer um *P* e um *T* respectivamente. Este pré-processamento facilita a utilização dos dados por parte das técnicas de obtenção dos índícios bem uma melhor aplicabilidade do método.

3.4.1 Extracção dos Lemas por categorias

A evolução das técnicas no domínio do PLN, fez com que os corpora no formato electrónico tenham evoluído de uma simples sequência de palavras para um estado rico de fonte linguística. Para além dos textos em si, os corpora podem conter um conjunto complexo de informações: informação semântica, sintáctica, morfológica, entre outros. Estes corpora

podem ser ditos como anotados, etiquetados ou enriquecidos. De entre esses, os corpora etiquetados morfologicamente são os mais comuns.

Sendo assim, nos corpora etiquetados morfologicamente, cada palavra do texto está associada a uma etiqueta morfológica que representa a sua categoria gramatical (substantivo, verbo, adjectivo, adverbio, entre outros).

A maioria dos trabalhos apresentam diferentes métodos de avaliação de indícios de subjectividade, no entanto, tendem a seleccionar, na maioria dos casos, as mesmas categorias morfológicas: adjectivos e verbos.

Sabendo que estas duas categorias são bons indícios de subjectividade, então vão ser tidas em conta. Mas como existem mais categorias morfológicas para além destas, as mesmas também vão ser tidas em conta, para verificar o seu comportamento.

Para obter a informação das categorias morfológicas e os lemas[8] das palavras presentes nos textos dos corpora, foi utilizado um etiquetador automático de categorias morfológicas. O etiquetador utilizado foi o *Treetagger*.

O Treetagger[9] [26] [27] é uma ferramenta utilizada para etiquetar o texto com as categorias morfológicas e a informação do lema. Foi desenvolvido por Helmut Schmid no projecto TC[10] no instituto de linguística computacional da Universidade de Stuttgart.

O Treetagger tem sido usado com sucesso na etiquetagem do alemão, inglês, francês, português, entre outros, e é adaptável a outros idiomas, se existir um léxico e um corpus de treino manualmente anotado.

Esta ferramenta utiliza métodos probabilísticos para efectuar a etiquetagem. Evita os problemas associados aos outros etiquetadores baseados no modelo de Markov, quando tem de calcular as probabilidades de transição a partir dos dados. Para isso, faz uso de árvores de decisão. Foi com base nisto que o Treetagger foi implementado, mostrando uma precisão de 96.36% na etiquetagem dos dados do Penn-Treebank, 0.3% melhor do que o trigram

[8] Forma básica da palavra, como ela aparece no dicionário.

[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[10] textual corpora and tools for their exploration (Textcorpora und Erschliessungswerkzeuge). http://www.ims.uni-stuttgart.de/projekte/tc/

tagger[11] com os mesmos dados.

A tabela 3.2 mostra um exemplo da forma como são etiquetados os dados, para a frase *o meu tio é português*.

Token	**Etiqueta**	**Lema**
o	DET	o
meu	ADJ	meu
tio	NOM	tio
é	V	ser
português	ADJ	português

Tabela 3.2: Exemplo de etiquetagem do Treetagger.

Após a aplicação do *Treetagger* aos corpora, foram construídos uns ficheiros com base na informação morfológica das palavras presentes nos textos. Os textos foram decompostos de acordo com a informação morfológica das palavras, isto é, por exemplo, o adjectivo *meu* foi guardado no ficheiro dos adjectivos juntamente com a informação do seu número de ocorrências nos textos bem como com o número de domínios em que aparece, isso no caso dos textos dos blogues.

Cada categoria morfológica é representada por dois ficheiros, o das palavras dos blogues (subjectivas) e o das palavras do CETEMPúblico (objectivas).

Palavra	**Frequência**	
seu	18155 (*)	
Palavra	**Domínios**	**Frequência**
meu	18	34236 (**)

Tabela 3.3: Um exemplo de uma palavra guardada nos respectivos ficheiros. (*) ficheiro objectivo; (**) ficheiro subjectivo

[11] Trigram tagger, etiquetador morfológico automático.

3.4.2 Extracção das palavras compostas por categorias

As palavras compostas são agrupamentos de palavras consecutivas compostas por um número fixo de palavras que, quando aparecem juntas no texto, tendem a ser subjectivas (Wiebe et al. [35]). Também existem agrupamentos de palavras não contíguas, no entanto estes ficam de fora do campo de estudo deste trabalho, podendo vir a ser utilizados em trabalhos futuros.

Nos estudos efectuados para a aprendizagem da linguagem subjectiva, este é outro tipo de indício de subjectividade com alta precisão. Existem dois tipos de agrupamentos de palavras contíguas, o primeiro caso é o que foi apresentado no inicio desta secção, composto por um número fixo de palavras; o segundo caso é igual com uma diferença que consiste na substituição das palavras que aparecem só uma vez nos dados e que fazem parte do agrupamento, pela palavra *UNIQUE* ou *U*.

Um *n-gram* contíguo de uma Unidade Textual (UT) é uma sequência ordenada de *n* unidades textuais correspondentes a uma sequência contínua de um texto sequencial não interrompido. A ordem da sequência é definida pela ordem em que aparecem as UT´s no texto (Dias [8]).

Neste trabalho só vai ser considerado o primeiro caso, as palavras compostas por um número fixo de palavras, podendo essas ser compostas por duas ou mais palavras, num máximo de cinco palavras consecutivas (*5-grams*).

Para obter as palavras compostas existentes nos textos dos corpora, foi utilizado um identificador automático de palavras compostas. O identificador utilizado foi o *SENTA*.

O software SENTA foi desenvolvido por Dias [8], e permite obter um conjunto de *n-grams* ou palavras compostas a partir do texto.

Palavras como *Vila Nova de Gaia, Sport Lisboa e Benfica, Presidente da Republica, tudo bem*, ocorrem frequentemente nos documentos, e utilizando o software, podem ser extraídas automaticamente, assumindo um significado próprio quando comparado com a sua forma singular.

O software SENTA utiliza métodos puramente estatísticos, baseados no número de vezes

que cada *n-gram* ocorre no corpus, utiliza o algoritmo GenLocalMaxs (Silva et al. [28]) e a medida Mutual Expectation (ME) (Dias et al. [9]), baseando-se na procura do máximo local da função de associação. Esta função mede a força da ligação existente entre vários *tokens*[12] de um *n-gram*.

As sequências de *tokens*, contíguas, fortemente ligadas entre si, correspondem a valores de ME elevados e serão escolhidos pelo algoritmo GenLocalMaxs como phrase[13].

A implementação do software baseou-se em particular no *suffix-array* (Gil e Dias [12]) de forma a extrair os *n-grams* em tempo real.

A tabela 3.4 mostra um exemplo da forma como são os dados dos n-grams extraídos pelo SENTA.

0.00138524349313	00019	Angra do Heroísmo
0.00102070579305	00014	Nossa Senhora
0.00093738280702	00015	Judas Tadeu
0.00036453775829	00005	base de dados

Tabela 3.4: Exemplo de n-grams extraídos pelo SENTA.

Como também queríamos saber as categorias morfológicas das palavras que fazem parte das palavras compostas, procurou-se os indícios das mesmas nos textos processados pelo etiquetador automático de categorias morfológicas.

Depois de obter a informação das categorias morfológicas também foi efectuada a decomposição dos textos em ficheiros de acordo com a informação morfológica, do mesmo modo que foi referido na secção 3.4, só que neste caso foi para as palavras compostas.

[12]Unidades textuais.

[13]Sequências contíguas de palavras.

Capítulo 4

Construção do Dicionário

Depois de ter apresentado no capítulo 3 os corpora bem como as operações computacionais a que eles foram submetidos, já se pode aplicar as técnicas estatísticas aos ficheiros objectivos e subjectivos de cada categoria morfológica dos dois tipos de indícios de subjectividade.

Estas técnicas estatísticas efectuam a filtragem das palavras, isto é, nem todo o universo de palavras contidas nos ficheiros subjectivos são subjectivas, porque se isso fosse tão linear não haveria necessidade de procurar palavras que representem a subjectividade. Assim, devido a essa não linearidade das palavras é necessário efectuar uma selecção das palavras de tal forma a que as palavras que se venham a obter sejam as mais subjectivas nos corpora em estudo.

Neste capítulo são apresentadas as diferentes etapas de filtragem dos ficheiros resultantes da computação dos corpora. Cada etapa foi aplicada em ambos as unidades lexicais da mesma maneira.

4.1 Verticalidade

O que se pretende com a aplicação deste filtro é efectuar uma separação entre as palavras objectivas e as palavras subjectivas que aparecem nos ficheiros para ambas as unidades lexicais. A verticalidade refere-se a essa separação, barreira física, construída com a equação 4.2, para filtrar as palavras.

A equação 4.2 é aplicada aos ficheiros de palavras (separados por categorias morfológicas) objectivos e subjectivos tal como se vai explicar de seguida. Para cada palavra dos ficheiros subjectivos calcula-se o seu peso com a equação 4.2, se este for superior a zero significa que a palavra é subjectiva porque aparece mais vezes no respectivo ficheiro subjectivo do que no respectivo ficheiro objectivo. Depois de ter calculado todos os pesos e depois de se ter removido as palavras com peso inferior e igual a zero, efectua-se o cálculo da média e todas as palavras com peso igual ou superior à média são seleccionadas.

Para calcular o peso utiliza-se uma adaptação para este problema da medida TF/IDF wieght (Term Frequency - Inverse Document Frequency).

$$W(i) = tf_i * log_2(\frac{N}{n_i}) \tag{4.1}$$

Efectuou-se a adaptação da fórmula porque de facto o peso que se pretende calcular não é para uma única palavra mas para ficheiros subjectivos e objectivos. Assim, a equação 4.2 serve para efectuar a selecção dessas palavras:

$$peso(i) = tf_{sub(i)} * log_{10}(\frac{tf_{sub(i)}}{tf_{obj(i)}}) \tag{4.2}$$

Onde, $peso(i)$ - Representa o peso de subjectividade da palavra i ao nível dos corpora; $tf_{sub(i)}$ - Representa a proporção da palavra i no corpus subjectivo; $tf_{obj(i)}$ - Representa a proporção da palavra i no corpus objectivo.

$$tf_{sub(i)} = \frac{N_{sub}}{N_{subTotal}} \tag{4.3}$$

Onde, $tf_{sub(i)}$ - Ver fórmula 4.2; N_{sub} - Número de vezes que a palavra i aparece no corpus subjectivo; $N_{subTotal}$ - Número total de palavras subjectivas do corpus subjectivo.

$$tf_{obj(i)} = \frac{N_{obj}}{N_{objTotal}} \tag{4.4}$$

Onde, $tf_{obj(i)}$ - Ver fórmula 4.2; N_{obj} - Número de vezes que a palavra i aparece no corpus objectivo; $N_{objTotal}$ - Número total de palavras subjectivas do corpus objectivo.

Depois de se ter aplicado esta fórmula obtemos uma lista de palavras ordenada por grau de relevância de subjectividade, isto para cada categoria. Depois de ter calculado todos os

pesos e depois de se ter removido as palavras com peso inferior e igual a zero, efectua-se o cálculo da média e todas as palavras com peso igual ou superior à média são seleccionadas.

A tabela 4.1 ilustra um exemplo da selecção das palavras subjectivas, onde *Freq. objectiva* significa a frequência de ocorrência da palavra no corpus objectivo; *Freq. subjectiva* significa a frequência de ocorrência da palavra no corpus subjectivo; *Selec.* significa se a palavra foi seleccionada.

Palavra + Categoria	**Freq. objectiva**	**Freq. subjectiva**	**Peso**	**Selec.**
seu_ADJ	60903	34909	0.1541	Sim
visivelmente_ADV	62	43	-0.00005	Não
o_DET seu_ADJ	2012	2824	0.1924	Sim
os_DET seus_ADJ	5176	926	-0.00682	Não

Tabela 4.1: Exemplo da escolha das palavras subjectivas.

4.2 Transversalidade

Após ter efectuado a separação entre as palavras objectivas e subjectivas, pretende-se identificar as palavras transversais aos vários domínios de subjectividade. Por transversalidade quer-se dizer que a palavra aparece em diferentes blogues da lista de blogues apresentada no apêndice A.

A equação 4.5 é aplicada aos ficheiros de palavras, separadas por categorias morfológicas, resultantes da etapa anterior (secção 4.1). Para cada palavra é calculado a sua importância, sendo que esta vai ser influenciada pelo número de blogues em que ela aparece vezes o logaritmo da frequência de ocorrência da palavra no corpus subjectivo. Depois de ter calculado todas as importâncias e depois de se ter removido as palavras com importância inferior e igual a zero, efectua-se o cálculo da média e todas as palavras com importância igual ou superior a média são seleccionadas.

$$importancia(i) = peso(i) * log_{10}(D * log_{10}(Freq(i))) \tag{4.5}$$

Onde, $importancia(i)$ -Representa a importância da palavra i ao nível dos blogues; $peso(i)$ - Ver fórmula 4.2; D - Número de blogues em que a palavra i aparece; $Freq(i)$ - Número de vezes que a palavra i aparece no corpus subjectivo.

Depois de se ter aplicado esta fórmula obtêm-se uma lista de palavras subjectivas ordenada por grau de relevância de domínios da subjectividade, isto para cada categoria.

4.3 Co-ocorrências

Após efectuar a selecção das palavras transversais aos domínios quer-se calcular a co-ocorrência das palavras subjectivas. A co-ocorrência de palavras diz respeito à possibilidade de palavras ocorrerem umas em combinação com outras.

Turney [31] apresentou uma propriedade especial na análise da polaridade, que se refere ao facto de que palavras com orientação similar tendem a co-ocorrerem. Como tal utilizou os dois termos *excellent* e *poor* como conjunto de palavras semente para determinar a orientação semântica de outras palavras. Este tipo de palavras semente pode ser vistos como um conector para as classes positiva e negativa respectivamente. Assim, termos que tendem a co-ocorrerem com a palavra *excellent* ao nível do documento tendem a ser positivas, e vice-versa com a palavra *poor*.

Assim, derivando esta propriedade apresentada por Turney [31] pretende-se calcular as co-ocorrências das palavras ao nível do documento, da frase e do contexto sobre os ficheiros obtidos na secção 4.2.

A derivação consiste no facto de que se pretende é dar mais valor às palavras obtidas até agora. Ou seja, na etapa da Verticalidade efectuou-se a separação entre as palavras objectivas e subjectivas. Depois na etapa da Transversalidade procuraram-se palavras subjectivas transversais aos vários domínios da subjectividade pertencentes a lista de blogues. Agora o que se pretende é obter as palavras mais subjectivas, isto é, o conjunto de palavras obtidas até agora são palavras subjectivas, então se elas co-ocorrem ao nível do documento (em primeiro), da frase (em segundo) e do contexto (em terceiro), significa que essas palavras são muito subjectivas por co-ocorrerem com outras palavras subjectivas.

Para cada etapa do cálculo da co-ocorrência, o valor da co-ocorrência é calculado de acordo com as respectivas equações com base nos resultados apresentados na matriz de co-ocorrência, onde cada célula contém o número de documentos, frases ou contexto, dependendo da co-ocorrência que esta a calcular.

A medida utilizada para calcular a co-ocorrência é SCP (Symmetric Conditional Probability),

$$SCP(P1, P2) = \frac{f(P1, P2)^2}{f(P1) * f(P2)} \tag{4.6}$$

A tabela 4.2 apresenta um exemplo da matriz de co-ocorrência,

Palavras	**P1**	**P2**	**P3**	...	**Pn**
P1		2	0	...	7
P2	3		6	...	0
P3	1	3		...	1
...	...	...	...		...
Pn	4	5	1	...	

Tabela 4.2: Matriz de co-ocorrência.

As palavras que têm um valor de co-ocorrência igual a zero não passam para o cálculo da próxima co-ocorrência, porque se o valor é zero significa que não co-ocorre com mais nenhuma das palavras provenientes dos ficheiros do cálculo da transversalidade.

Depois de calcular o valor dessas co-ocorrências, os valores obtidos são normalizados e efectua-se a soma de cada componente das co-ocorrências, para cada palavra, depois efectua-se o cálculo da média e todas as palavras com soma de co-ocorrências igual ou superior a média são seleccionadas, obtendo uma lista de palavras ordenadas pela soma das co-ocorrências, isto para cada categoria.

4.3.1 A nível do documento

A equação 4.7 serve para calcular a co-ocorrência das palavras subjectivas ao nível do documento.

$$scpDoc(x, y) = \frac{f(x, y)^2}{f(x) * f(y)} \tag{4.7}$$

Onde, $scpDoc(x, y)$ - Representa a co-ocorrência da palavra subjectiva x com a palavra subjectiva y, ao nível do documento; $f(x, y)$ - Número de documentos em que co-ocorrem as palavras subjectivas x e y; $f(i)$ - Número de documentos em que a palavra subjectiva i aparece, onde $i = x; y$.

Depois de aplicar esta fórmula faz-se a soma das co-ocorrências de cada palavra, as palavras com maior soma devem ser mais subjectivas porque andam nos mesmos documentos. Obtêm-se uma lista de palavras subjectivas ordenada por co-ocorrência ao nível do documento, isto para cada categoria.

4.3.2 A nível da frase

A equação 4.8 serve para calcular a co-ocorrência das palavras subjectivas ao nível da frase.

$$scpFrase(x, y) = \frac{f(x, y)^2}{f(x) * f(y)} \tag{4.8}$$

Onde, $scpFrase(x, y)$ - Representa a co-ocorrência da palavra subjectiva x com a palavra subjectiva y, ao nível da frase; $f(x, y)$ - Número de contextos em que co-ocorrem as palavras subjectivas x e y; $f(i)$ - Número de contextos em que a palavra subjectiva i aparece, onde $i = x; y$.

Depois de aplicar esta fórmula faz-se a soma das co-ocorrências de cada palavra, as palavras com maior soma devem ser mais subjectivas porque andam nas mesmas frases. Obtém-se uma lista de palavras subjectivas ordenada por co-ocorrência ao nível da frase, isto para cada categoria.

4.3.3 A nível do contexto

A equação 4.9 serve para calcular a co-ocorrência das palavras subjectivas ao nível do contexto.

$$scpContexto(x, y) = \frac{f(x, y)^2}{f(x) * f(y)} \tag{4.9}$$

Onde, $scpContexto(x, y)$ - Representa a co-ocorrência da palavra subjectiva x com a

palavra subjectiva y, ao nível do contexto; $f(x, y)$ - Número de contextos em que co-ocorrem as palavras subjectivas x e y; $f(i)$ - Número de contextos em que a palavra subjectiva i aparece, onde $i = x; y$.

Depois de aplicar esta fórmula faz-se a soma das co-ocorrências de cada palavra, as palavras com maior soma devem ser mais subjectivas porque andam nos mesmos contextos. Obtêm-se uma lista de palavras subjectivas ordenada por co-ocorrência ao nível do contexto, isto para cada categoria.

Capítulo 5

Resultados

Após ter aplicado as técnicas estatísticas para efectuar a filtragem das palavras, obteve-se um conjunto de palavras e expressões (n-grams) que fazem parte do léxico subjectivo construído automaticamente para o português europeu (ver apêndice B e C).

Para avaliar este léxico subjectivo efectuou-se uma análise dos dados direccionada para a análise qualitativa e a análise quantitativa. Estas análises foram efectuadas com o auxilio de uma plataforma de validação desenvolvida para esta tese.

Figura 5.1: Plataforma de validção do léxico subjectivo.

5.1 Análise qualitativa

As palavras e expressões subjectivas obtidas após a filtragem efectuada com técnicas estatísticas (ver capítulo 4), permitem efectuar uma análise qualitativa. Para isso é necessário que existam certas qualidades esperadas.

Essas qualidades dizem respeito às palavras e expressões subjectivas. Contudo como foi referido na secção 2.1 o sentimento é sinónimo de subjectividade, mas o sentimento não é só subjectividade, também é subjectividade polar. Isto é, as palavras subjectivas podem ter polaridade inerente.

Nesta análise não há necessidade de se saber quantas palavras e expressões são subjectivas ou subjectivas polares, apenas precisa-se de comparar a qualidade observada com os padrões pré-estabelecidos.

Assim, de seguida as tabelas apresentam a análise qualitativa efectuada sobre as palavras e expressões do léxico subjectivo, apresentadas nos apêndices B e C.

As tabelas 5.1 e 5.2 apresentam as palavras e expressões subjectivas polares, pertencentes ao léxico subjectivo.

Categoria morfológica	**Palavras**
Adjectivo	bom; melhor
Adverbio	não; bem
Nome	nem; não; heroísmo; nada; melhor; bem; sempre; problema; mal; problemas; alegria; amor; amigo; certeza; agradecimento; realidade; humildade; saúde;

Tabela 5.1: Palavras subjectivas polares.

Expressões	Expressões
bom_ADJ gosto_NOM	nem_CONJ se_P
muito_ADV mais_ADV	ir_V muito_ADV
muito_ADV bem_ADV	ter_V mais_ADV
bom_ADJ senso_NOM	ser_V muito_ADV
muita_ADJ gente_NOM	ser_V mais_ADV
bem_ADV mais_ADV	muito_ADV mal_ADV

Tabela 5.2: Expressões subjectivas polares.

As tabelas 5.3 e 5.4 apresentam as palavras e expressões subjectivas, pertencentes ao léxico subjectivo.

Categorias morfológicas	Palavras
Adjectivo	seu; meu; sua; minha; qualquer; grande; mellhor; nossa; nosso; mesmo; íntimo; maior; teu; alguma; muita; caro; novo; tua; algum
Adeverbio	já; mais; ainda; muito; só; agora
Nome	mas; até; vez; me; mais; tão; vezes; mesmo; nossa; muito; ainda; eu; meu; sua; nós; razão; orgulho; maioria; espírito; opinião; nosso; poder; comentários; ele; grande; minha; vontade
Pronome	o; me; eu
Preposição + Pronome	nesta; desta; deste
Verbos	ver; fazer; dar; dizer; deixar; ir; chamar; ficar;ser
Verbos + Pronome	dar-lhes; atrevo-me; dá-nos; fez-me; recordo-me; faz-me; leva-me; senti-me; disse-lhe; dizer-lhe; dou-lhe; deixa-me; refiro-me; foi-me; dar-lhe;faz-nos

Tabela 5.3: Palavras subjectivas.

Expressões
todo_ADJ a_DET
seu_ADJ marido_NOM
meu_ADJ marido_NOM
minha_ADJ filha_NOM
seu_ADJ filho_NOM
meu_ADJ filho_NOM
seu_ADJ pais_NOM
meu_ADJ pai_NOM

Expressões
meu_ADJ amigo_NOM
quase_ADV sempre_ADV
algum_ADJ tempo_NOM
que_CONJSUB lhe_P
nossa_ADJ senhora_NOM
ponto_NOM de_PRP vista_NOM
ser_V apenas_ADV

Tabela 5.4: Expressões subjectivas.

5.2 Análise quantitativa

Após ter efectuado a análise qualitativa pode-se efectuar uma análise quantitativa, porque a secção anterior forneceu elementos adicionais que permitem efectuar este tipo de análise.

Como já foi referido, os sentimentos podem ser subjectivos e podem ser subjectivos polares. Como tal quer-se saber qual a precisão da construção do léxico subjectivo para cada tipo de palavra pertencente às distintas categorias morfológicas e para as expressões.

Para calcular a precisão utilizou-se a equação 5.1 e os dados das tabelas 5.1, 5.2, 5.3, 5.4, 5.5 e 5.6.

$$precisao(i) = \frac{\sum_{j=1}^{2} C_j}{NP_i} \tag{5.1}$$

Onde, $precisao(i)$ - Representa a precisão da categoria morfológica ou expressões i; C_j - Número de palavras ou expressões subjectivas, subjectiva polares; NP_i - Número total de palavras emocionais obtidas para a categoria morfológica ou expressões i.

As tabelas 5.5 e 5.6 apresentam o número total de palavras por categorias morfológicas e o número total de expressões contidas no léxico emocional.

Categorias morfológicas	Número palavras
Adjectivo	23
Adverbio	8
Determinante	1
Nome	199
Pronome	3
Preposição + Pronome	3
Verbo	9
Verbo + Pronome	20
Total	**266**

Tabela 5.5: Número total de palavras por categoria morfológica.

n-grams	Número palavras
Total	211

Tabela 5.6: Número total de expressões.

As tabelas 5.7 e 5.8 apresentam a precisão das expressões e as precisões para cada categoria morfológica contidas no léxico emocional.

n-grams		
	Polar	12
	Subjectivo	15
	Precisão	12.8%

Tabela 5.7: Precisão das expressões.

Adjectivo	**Polar**	2	**Pronome**	**Polar**	0
	Subjectivo	19		**Subjectivo**	3
	Precisão	91.3%		**Precisão**	100%
Adverbio	**Polar**	2	**Pronome + Preposição**	**Polar**	0
	Subjectivo	6		**Subjectivo**	3
	Precisão	100%		**Precisão**	100%
Determinante	**Polar**	0	**Verbo**	**Polar**	0
	Subjectivo	0		**Subjectivo**	9
	Precisão	0%		**Precisão**	100%
Nome	**Polar**	18	**Verbo + Pronome**	**Polar**	0
	Subjectivo	27		**Subjectivo**	16
	Precisão	22.61%		**Precisão**	80%

Tabela 5.8: Precisão de cada categoria morfológica.

Da análise da tabela 5.8 verifica-se que as precisões de cada categoria morfológica apresentam em grande parte os resultados esperados. Contudo há uma categoria morfológica que não apresenta a precisão que se esperava, essa categoria morfológica é a dos nomes.

Ao reanalisar a tabela da categoria morfológica dos nomes que aparece no apêndice B e as tabelas 5.1 e 5.3 que apresentam as palavras subjectivas polares e as palavras subjectivas, verifica-se que apesar da grande quantidade de palavras obtidas, estas são demasiado tópicas. Ou seja, na maioria dos casos as palavras não apresentam qualquer tipo de indício de subjectividade. Elas podem ser subjectivas é nos domínios específicos onde elas aparecem em conjugação com outras palavras ou expressões.

O facto da categoria morfológica dos determinantes apresentar uma precisão igual a zero não é muito surpreendente porque só existe uma palavra e os determinantes não costumam ser indícios de subjectividade, daí não se estranhar muito este resultado.

Ao analisar a tabela 5.7 verifica-se que a precisão dos n-grams (expressões), não é a que se esperava. Como tal efectuou-se o mesmo processo do que no caso dos nomes, mas desta vez para as expressões, e verificou-se que o problema apresentado na categoria morfológica dos nomes era ainda mais notório neste tipo de indício de subjectividade.

Capítulo 6

Conclusões e Trabalho Futuro

6.1 Conclusões

No inicio de um novo projecto cientifico é necessário criar uma base sólida de conhecimentos. Neste sentido e para o desenvolvimento desta tese, estudou-se ao detalhe os estudos existentes na área da análise de sentimento que construíram ou utilizaram um dicionário emocional. Com isto aprofundaram-se os conhecimentos e avaliaram-se as novas soluções possíveis de proposta.

Deste modo, nesta tese foi proposto um novo método de construção automática de dicionários emocionais (dicionários subjectivos) para a língua portuguesa europeia, utilizando técnicas estatísticas e corpora não anotados manualmente.

Este novo método fornece à comunidade cientifica um outro caminho de investigação na área da análise de subjectividade.

O utilizador em geral, também é beneficiado, pois estes dicionários podem servir de base para implementar classificadores de subjectividade que avaliam os conteúdos das páginas Web. Assim, o utilizador sabe com algum grau de certeza se a informação que visualiza é verídica ou se é apenas um conjunto de opiniões/sentimentos expressos pelos autores, excluindo, desta forma os seus próprios juízos de valor relativamente ao que está a visualizar.

Os corpora foram construídos com base nos resultados apresentados por Pais [21], na avaliação da similaridade entre corpora manualmente anotados e corpora não anotados manualmente. Esses resultados mostraram que se pode considerar um corpus jornalístico

como sendo um corpus objectivo e que um conjunto de blogues pode ser tido em conta como um corpus subjectivo (ver secção 2.5).

Assim, utilizou-se o corpus jornalístico CETEMPúblico como corpus objectivo e um conjunto de blogues como corpus subjectivo.

Após ter descrito os corpora na secção 3.3, efectuou-se a extracção de forma automática dos indícios de subjectividade, que segundo os estudos existentes na análise de subjectividade, apresentam melhor a informação subjectiva nos documentos, textos e frases. Estes indícios são apresentados na secção 3.4 e dizem respeito as categorias morfológicas da palavras simples (part-of-speech) e as palavras compostas (sequencias de palavras contíguas).

A aplicação das técnicas estatísticas aos indícios de subjectividade, para construir o dicionário emocional, apresenta bons resultados em grande parte das categorias morfológicas. No entanto existem uma categoria (os nomes) que não apresenta os resultados esperados, no sentido que as palavras resultantes serem palavras tópicas, não subjectivas. Ou seja, na maioria dos casos as palavras não apresentam qualquer tipo de indício de subjectividade, elas podem ser subjectivas é nos domínios específicos onde elas aparecem em conjugação com outras palavras ou expressões.

Isso sucede no caso das palavras simples pertencentes à categoria morfológica dos nomes bem como no caso das palavras compostas, sendo que nestas últimos os resultados obtidos apresentam muito mais palavras tópicas do que no caso das palavras simples.

Com este trabalho demonstrou-se que é possível construir automaticamente um léxico de subjectividade para a língua portuguesa europeia, aplicando técnicas estatísticas e utilizando corpora não anotados manualmente e ferramentas para extrair automáticamente os indícios de subjectividade. Com a construção deste léxico de subjectividade é possível testar e/ou desenvolver novos classificadores de subjectividade para o português. Desta forma o conjunto que é composto por recursos semelhante é enriquecido e aumentado com um novo léxico para outra língua que não o inglês. Resta salientar que este léxico de subjectividade, construído automáticamente, constituí a primeira análise de subjectividade para o português europeu (depois de muita pesquisa, verificou-se que não existe um recurso semelhante para a língua portuguesa europeia).

6.2 Trabalho Futuro

Para realizar este trabalho de investigação científica, definiu-se um plano de trabalho que não a acaba com a conclusão desta tese de mestrado. Existem várias linhas de desenvolvimento futuras que podem ser abordadas na sequência dos resultados obtidos e dos trabalhos experimentais apresentados nesta tese. Apesar dos resultados obtidos, este trabalho pode ser muito mais desenvolvido e enriquecido.

De seguida são definidos um conjunto de ideias para melhorar os resultados obtidos e atingir com sucesso o objectivo traçado para esta tese.

A primeira sugestão consiste em aplicar um método de clustering[1] com o algoritmo Stuctural Correspondence Learning (SCL) [3][4]. Como já foi referido na secção 6.1 foram obtidos bons resultados para grande parte das categorias morfológicas, no entanto, nos nomes os resultados obtidos consistem em palavras de tópicos e não em palavras subjectivas. Assim, no sentido de remover as palavras por tópico sugere-se aplicar um método de clustering com o algoritmo SCL.

A ideia chave do SCL é identificar as correspondências entre as características dos diferentes domínios modelando as suas correlações com as características pivô. As características pivô são características que se comportam da mesma forma para a aprendizagem discriminativa em ambos os domínios.

Supondo que se pretende adaptar comentários de computadores a comentários de telemóveis. Enquanto muitas das características dos comentários de telemóveis e dos computadores são as mesmas, por exemplo *excelente* e *horrível*, muitas das palavras são completamente novas, como por exemplo *recepção*. Ao mesmo tempo muitas das características úteis para os computadores, tal como *dual-core* não são úteis para os telemóveis. A principal intuição é que mesmo quando boa *qualidade de recepção* e *dual-core rápido* são completamente distintas para cada domínio, se ambas tem alta correlação com *excelente* e baixa correlação com *horrível*, sobre os dados marcados, então pode-se tentar alinhá-los.

Depois de efectuar a aprendizagem de um classificador para comentários de computadores, quando se vêem características de telemóveis como *boa qualidade de recepção* sabe-se que ele deve comportar-se de uma forma semelhante que *dual-core rápido*.

[1] O clustering É uma técnica de Data Mining para fazer agrupamentos automáticos de dados segundo o seu grau de semelança.

A primeira etapa do SCL [3] é definir um conjunto de características pivô sobre os dados não etiquetados ao nível morfológico a partir de ambos os domínios. A selecção desses pivô baseia-se nas frequências de ocorrência bem como em função da sua informação mútua com as etiquetas de origem.

Depois utiliza-se um classificador treinado num dado domínio sobre uns dados de outro domínio diferente para efectuar a adaptação dos diferentes domínios.

A segunda sugestão consiste em efectuar a marcação das palavras subjectivas obtidas, em termos da sua polaridade. Em grande parte dos trabalhos experimentais apresentados, a análise dos conteúdos é direccionada no sentido de efectuar a análise de sentimento expresso nas opiniões em termos da sua polaridade (positiva, negativa o neutra).

A terceira sugestão consiste na implementação dos classificadores para efectuar as respectivas classificações de sentimentos, ou seja, implementar o classificador de subjectividade, para realizar a classificação subjectiva dos sentimentos; e implementar o classificador de polaridade, para realizar a classificação polar dos sentimentos. Ambos os classificadores são implementados para efectuar as respectivas classificações dos sentimentos para o português europeu.

Apêndice A

Lista dos domínios dos blogues

- A invenção de morel extraiu-se 164 blogues:
  http://morel.weblog.com.pt/

- Abnoxio extraiu-se 45 blogues:
  http://abnoxio.weblog.com.pt/

- Água lisa extraiu-se 435 blogues:
  http://agualisa6.blogs.sapo.pt/

- Arrastão extraiu-se 621 blogues:
  http://arrastao.org/

- Azoriana extraiu-se 3028 blogues:
  http://silvarosamaria.blogs.sapo.pt/

- Budismo extraiu-se 264 blogues:
  http://budismo.blogs.sapo.pt/

- Chica Ilheu extraiu-se 3951 blogues:
  http://chicailheu.blogs.sapo.pt/

- Erga Omnes extraiu-se 2171 blogues:
  http://ergaomnes.blogs.sapo.pt/

- Eugenio Melo e Castro extraiu-se 128 blogues:
  http://poportugal.blogs.sapo.pt/

- Formiguinha atómica extraiu-se 2725 blogues:
  http://formiguinha.blogs.sapo.pt/

- Gasolim extraiu-se 2262 blogues:
  http://gasolim.blogs.sapo.pt/

- Harry Potter news extraiu-se 740 blogues:
  http://harrypotternews.blogs.sapo.pt/

- Mundial 2006 extraiu-se 9291 blogues:
  http://mundial2006.blogs.sapo.pt/

- Ondas3 extraiu-se 1995 blogues:
  http://ondas3.blogs.sapo.pt/

- Papiro extraiu-se 593 blogues:
  http://papiro.blogs.sapo.pt/

- Porto das pipas extraiu-se 964 blogues:
  http://portodaspipas.blogs.sapo.pt/

- Quadratura do círculo extraiu-se 1983 blogues:
  http://quadraturadocirculo.blogs.sapo.pt/

- Sonhos reais extraiu-se 54 blogues:
  http://gorety.blogs.sapo.pt/

- Trilogia da herança extraiu-se 91 blogues:
  http://inheritancetrilogy.blogs.sapo.pt/6393.html

Apêndice B

Listas de lemas subjectivos por categorias morfológicas

As palavras estão ordenadas por ordem de importância da soma das co-ocorrências normalizadas.

Palavra	**Domínios**	**Frequência**	**Soma**
seu	18	34909	2,1214
meu	18	34236	2,1012
sua	18	33399	1,9464
minha	18	25181	1,6988
qualquer	18	13654	1,53
grande	18	18801	1,4053
bom	18	27641	1,3772
melhor	18	14087	1,3159
cada	18	10798	1,2794
todo	18	42105	1,2704
nossa	18	14968	1,2456
outro	18	31518	1,2401

Palavra	Domínios	Frequência	Soma
nosso	18	14481	1,2356
mesmo	18	16768	1,1776
íntimo	9	140	1,0468
maior	17	8328	1,0358
teu	18	8264	0,9833
alguma	18	11131	0,9126
muita	18	11981	0,8297
caro	17	3223	0,8064
novo	18	14384	0,7867
tua	17	5681	0,7739
algum	18	12060	0,77

Tabela B.1: Adjectivos subjectivos.

Palavra	Domínios	Frequência	Soma
não	19	209383	2,1383
já	19	52685	1,8489
mais	19	87890	1,8232
ainda	19	29048	1,7009

Palavra	Domínios	Frequência	Soma
muito	19	42054	1,658
bem	19	34598	1,6167
só	19	32183	1,5408
agora	19	21257	1,4354

Tabela B.2: Adverbios subjectivos.

Palavra	Domínios	Frequência	Soma
a	18	461538	1,0292

Tabela B.3: Determinantes subjectivos.

Palavra	Domínios	Frequência	Soma
que	18	12784	1,9796
e	18	11057	1,6859
de	18	12489	1,6137
a	19	16194	1,466
nem	9	669	1,4086
o	16	3276	1,3959
com	15	3057	1,3868
não	18	7696	1,3098
mas	16	2809	1,2225
isso	12	1231	1,1982
um	17	4617	1,1845
por	16	3050	1,1796
é	11	1672	1,1795
tem	9	1334	1,1663
portugal	19	37107	1,1533
assim	8	709	1,1476
os	13	1204	1,1361
para	17	4063	1,1296
se	17	3186	1,1044
em	16	2970	1,0882
do	18	4873	1,0693
como	16	2178	1,0347
porque	12	1307	1,0286
varandas	11	524	1,0103
dia	19	26072	0,9841
vai	9	456	0,979
até	11	735	0,9521
vez	19	17888	0,9509
me	13	932	0,9499
só	14	1839	0,9428
no	17	2700	0,9282
mais	19	2208	0,9152
mundo	19	13710	0,9071
tempo	19	16926	0,9029
anos	19	21098	0,9003
heroísmo	9	1710	0,8972

Palavra	Domínios	Frequência	Soma
uma	16	2245	0,8927
dos	14	1330	0,8846
da	18	2772	0,8755
todos	18	1922	0,8589
vida	19	14914	0,8566
fazer	9	547	0,8543
ao	15	1348	0,8498
país	17	15271	0,843
pessoas	19	14569	0,8421
nos	15	1164	0,8297
este	11	646	0,8227
ser	19	7545	0,8153
forma	18	11535	0,8053
quem	15	663	0,7978
portugueses	18	11462	0,7957
tão	9	738	0,793
futebol	16	15501	0,7863
na	16	1720	0,7803
parte	19	11833	0,7758
angra	7	3289	0,7657
coisa	19	12819	0,7521
quando	9	516	0,7481
vezes	19	10179	0,7341
já	11	727	0,7266
selecção	14	15584	0,7202
casa	19	10358	0,7108
as	18	1866	0,7085
pelo	13	545	0,7053
dias	19	11585	0,6994
cristiano	7	1316	0,6916
aos	13	530	0,6875
nada	15	1174	0,6852
ano	19	11804	0,6792
coisas	19	9680	0,6714
mesmo	14	818	0,6704
fim	19	7344	0,6676

Palavra	Domínios	Frequência	Soma
estão	9	529	0,6607
trabalho	19	9099	0,6593
gente	19	7766	0,656
senhora	18	4348	0,6525
ilha	16	7184	0,6498
nossa	15	3453	0,6494
estado	17	10576	0,6449
scolari	11	11787	0,6416
lado	19	7864	0,636
muito	15	1672	0,635
ainda	16	910	0,6338
eu	18	2124	0,6298
falta	18	6723	0,6277
ronaldo	10	2241	0,6241
caso	18	8975	0,6228
aqui	13	869	0,6213
verdade	19	6390	0,6198
são	17	4647	0,619
jogadores	13	7707	0,6116
melhor	13	962	0,607
nas	14	569	0,6018
português	18	5208	0,5984
justiça	17	11850	0,5975
foi	15	949	0,5958
nome	19	6393	0,5934
terceira	11	4111	0,5896
facto	19	6787	0,5838
também	12	760	0,58
mundial	15	4819	0,5775
josé	17	5962	0,5761
brasil	17	8946	0,5736
meu	14	1203	0,5691
povo	15	6427	0,5683
bem	15	1247	0,5666
final	17	5381	0,5662
jogo	17	8252	0,5658

Palavra	Domínios	Frequência	Soma
bandeiras	15	1261	0,5573
sua	12	642	0,5572
pena	19	5526	0,5481
sempre	16	1180	0,5456
força	18	7106	0,5449
momento	18	5019	0,5447
exemplo	19	4981	0,5441
amigos	18	6408	0,5434
noite	19	6487	0,54
lugar	19	6321	0,5395
governo	17	10784	0,5394
horas	18	5770	0,5322
história	19	5840	0,5321
tipo	18	4553	0,5285
hora	18	4583	0,5279
altura	19	4558	0,5265
nós	15	878	0,5235
maria	16	5981	0,5205
joão	18	5941	0,518
terra	18	5118	0,5094
razão	18	4924	0,5086
sentido	18	4156	0,5078
milagres	16	1312	0,5036
problema	18	4941	0,5002
orgulho	16	3342	0,4999
pessoa	19	4531	0,4997
palavras	19	4905	0,4996
costa	18	5083	0,4985
mal	14	1602	0,4972
maioria	19	4468	0,4963
senhor	16	6020	0,4963
respeito	17	3787	0,496
espírito	18	2733	0,4959
semana	18	4791	0,4873
opinião	18	4754	0,484
nosso	9	728	0,4836

Palavra	Domínios	Frequência	Soma
portuguesa	17	2415	0,4823
olhos	19	3898	0,4813
pais	17	4418	0,4788
frente	18	4799	0,4787
poder	19	5679	0,4785
coração	17	5248	0,4773
problemas	17	3773	0,4767
dinheiro	18	5243	0,4723
paulo	17	4636	0,4717
deus	19	5729	0,4706
ministério	14	5815	0,4706
meio	18	3855	0,4688
filhos	19	3392	0,4688
ponto	18	3190	0,4629
comentários	17	4713	0,4623
santo	16	2314	0,4619
equipa	16	5096	0,4611
condições	17	3407	0,4605
ele	13	1024	0,4604
europa	17	3488	0,4603
causa	18	4341	0,4584
questão	18	4477	0,4578
the	16	4022	0,457
direito	18	4881	0,4547
and	14	2171	0,4525
silva	17	4700	0,4525
brasileiros	12	1867	0,4511
alegria	16	3989	0,4504
amor	18	5839	0,4468
amigo	18	4433	0,4461
porto	18	5989	0,4454
jogos	16	3780	0,4448
das	14	552	0,4421
certeza	17	3249	0,4392
ministro	16	5000	0,4379
relação	19	3562	0,4361

Palavra	Domínios	Frequência	Soma
cabeça	17	3635	0,4359
imagem	19	4428	0,4354
agradecimento	13	982	0,4331
césar	12	1346	0,4294
favor	19	3986	0,4293
açores	16	4100	0,428
grande	16	2017	0,4266
realidade	17	3200	0,4266
humildade	15	717	0,4255
meses	19	3876	0,425
vista	18	2596	0,4225
saúde	16	4391	0,4222
minha	12	936	0,4202
comentário	18	4507	0,4188
luís	16	3204	0,4174
vontade	19	3588	0,417
mar	18	4596	0,4149
serviço	17	3751	0,4145
casos	17	3048	0,4137

Tabela B.4: Nomes subjectivos.

Palavra	Domínios	Frequência	Soma
o	18	32623	0,1419
me	18	36069	0,1416
eu	18	47017	0,1381

Tabela B.5: Pronome subjectivos.

Palavra	Domínios	Frequência	Soma
nesta	18	5128	0,0601
desta	18	6764	0,058
deste	19	9207	0,0538

Tabela B.6: Preposiçao + Pronome subjectivos.

Palavra	Domínios	Frequência	Soma
ver	17	2677	0,1395
fazer	17	4576	0,1349
dar	18	3669	0,1325
dizer	17	3739	0,1302
deixar	18	2430	0,1259
ir	18	1695	0,1255
chamar	17	1319	0,1137
ficar	12	629	0,1134
ser	15	793	0,111

Tabela B.7: Verbos subjectivos.

Palavra	**Domínios**	**Frequência**	**Soma**
dar-lhes	10	183	0,0767
atrevo-me	9	95	0,0537
dá-nos	12	163	0,0245
fez-me	10	420	0,0238
recordo-me	7	150	0,0229
faz-me	14	503	0,0228
leva-me	9	129	0,0222
senti-me	9	177	0,0221
disse-lhe	11	190	0,0221
dizer-lhe	13	181	0,0197
dou-lhe	7	101	0,0191
diz-se	15	284	0,0184
deixa-me	12	148	0,0176
refiro-me	12	172	0,0173
fazer-se	12	214	0,0172
foi-me	11	145	0,0166
dar-lhe	13	299	0,0166
esquecem-se	8	136	0,0164
vê-la	12	154	0,0162
faz-nos	10	112	0,0161

Tabela B.8: Verbos + Pronome subjectivos.

Apêndice C

Listas de n-grams subjectivos

Palavra	**Domínios**	**Frequência**
todo_ADJ a_DET	13	919
milton_NOM nascimento_NOM	1	34
caetano_NOM veloso_NOM	1	73
ney_NOM matogrosso_NOM	1	14
seu_ADJ marido_NOM	1	23
meu_ADJ marido_NOM	2	496
que_CONJSUB a_DET	14	5316
que_CONJSUB o_DET	16	10693
minha_ADJ filha_NOM	3	125
seu_ADJ filho_NOM	1	16
um_DET projecto_NOM	1	72
meu_ADJ filho_NOM	5	147
de_PRP um_DET	13	9622
uma_DET campanha_NOM	1	20
com_PRP a_DET	17	8435
para_PRP o_DET	15	13569
de_PRP uma_DET	11	4425
com_PRP o_DET	18	12131
para_PRP a_DET	16	8218
bom_ADJ gosto_NOM	1	33

Palavra	**Domínios**	**Frequência**
chico_NOM buarque_NOM	1	277
seu_ADJ pais_NOM	1	15
uma_DET mulher_NOM	4	82
vera_NOM jardim_NOM	1	36
como_CONJ o_DET	10	239
tribunal_NOM administrativo_NOM	1	40
como_CONJ um_DET	3	52
como_CONJ a_DET	3	47
um_DET dia_NOM	9	608
muito_ADV mais_ADV	11	254
um_DET homem_NOM	1	15
comunidade_NOM internacional_ADJ	1	12
meu_ADJ pai_NOM	4	189
luz_NOM verde_ADJ	1	15
muito_ADV bem_ADV	12	526
com_PRP um_DET	12	990
alguma_ADJ coisa_NOM	12	277
por_PRP uma_DET	6	144
por_PRP um_DET	9	316
com_PRP uma_DET	10	646
um_DET problema_NOM	2	11
se_CONJ o_DET	2	10
um_DET livro_NOM	5	71
meu_ADJ amigo_NOM	4	54
sobre_PRP a_DET	12	291
supremo_NOM tribunal_NOM	1	58
ferro_NOM rodrigues_NOM	2	84
sobre_PRP o_DET	13	536
parque_NOM natural_ADJ	1	42
uma_DET pessoa_NOM	7	277
quase_ADV sempre_ADV	1	23
como_CONJ uma_DET	5	158
um_DET processo_NOM	1	86
bom_ADJ senso_NOM	6	59
sociedade_NOM civil_ADJ	1	23
nobre_NOM guedes_NOM	1	13

Palavra	**Domínios**	**Frequência**
para_PRP uma_DET	2	33
uma_DET noite_NOM	1	18
contra_PRP a_DET	3	86
entre_PRP a_DET	4	37
para_PRP um_DET	4	71
muita_ADJ gente_NOM	11	178
uma_DET empresa_NOM	2	20
que_CONJSUB o_P	14	685
este_DET ano_NOM	8	200
muito_ADV mal_ADV	2	8
por_PRP este_DET	3	24
este_DET sistema_NOM	1	18
algum_ADJ tempo_NOM	5	48
ana_NOM benavente_NOM	1	48
contra_PRP o_DET	7	75
entre_PRP o_DET	8	162
que_CONJSUB lhe_P	2	8
bem_ADV mais_ADV	3	65
que_CONJSUB a_P	12	176
campanha_NOM eleitoral_ADJ	2	65
durante_PRP a_DET	1	19
programa_NOM eleitoral_ADJ	1	14
uma_DET equipa_NOM	2	85
um_DET trabalho_NOM	4	34
que_CONJSUB se_P	17	4613
um_DET filme_NOM	3	37
uma_DET proposta_NOM	2	14
ana_NOM maria_NOM	1	12
fernando_NOM pessoa_NOM	4	38
que_PR a_P	14	690
vitalino_NOM canas_NOM	1	44
este_DET caso_NOM	1	12
uma_DET prova_NOM	1	33
que_PR lhe_P	6	57
que_CONJSUB este_P	2	5
santana_NOM lopes_NOM	4	288

Palavra	**Domínios**	**Frequência**
uma_DET imagem_NOM	2	32
que_PR o_P	14	2295
manuel_NOM alegre_NOM	6	215
que_PR ela_P	4	98
quem_PR lhe_P	1	16
castelo_NOM branco_NOM	2	91
telmo_NOM correia_NOM	2	24
partido_NOM socialista_NOM	2	51
guerra_NOM civil_ADJ	1	48
florbela_NOM espanca_NOM	1	11
que_PR ele_P	7	488
paulo_NOM portas_NOM	4	168
um_DET jogo_NOM	1	241
ir_V a_DET	2	10
pacheco_NOM pereira_V	1	16
cavaco_NOM silva_NOM	4	322
jorge_NOM coelho_NOM	2	101
marques_NOM mendes_NOM	6	203
para_PRP que_PR	7	238
de_PRP que_PR	7	1446
como_CONJ ele_P	2	14
que_CONJSUB ela_P	3	50
por_PRP cento_NOM	4	18
primeiro_NOM ministro_NOM	4	30
jorge_NOM sampaio_NOM	2	134
branqueamento_NOM de_PRP capitais_NOM	1	38
new_NOM york_NOM	1	45
estados_NOM membros_NOM	1	29
nossa_ADJ senhora_NOM	3	87
em_PRP que_PR	14	2786
parlamento_NOM europeu_NOM	2	48
of_NOM the_NOM	3	95
campo_NOM de_PRP golfe_NOM	1	20
que_PR a_DET	15	12708
tribunal_NOM constitucional_NOM	2	45
reino_NOM unido_NOM	3	219

Palavra	**Domínios**	**Frequência**
george_NOM bush_NOM	1	11
como_CONJ o_P	2	28
um_DET jogador_NOM	1	79
nem_CONJ se_P	2	62
consumo_NOM de_PRP energia_NOM	1	13
efeito_NOM de_PRP estufa_NOM	1	16
pacheco_NOM pereira_NOM	3	278
tony_NOM blair_NOM	2	30
jorge_NOM costa_NOM	1	69
por_PRP este_P	3	11
proposta_NOM de_PRP lei_NOM	1	74
nova_NOM iorque_NOM	4	32
ser_V a_DET	14	1364
que_PR o_DET	16	27053
estados_NOM unidos_NOM	8	174
ter_V uma_DET	6	87
assembleia_NOM municipal_NOM	2	46
que_PR se_P	18	11166
ir_V muito_ADV	1	36
projecto_NOM de_PRP lei_NOM	1	12
bases_NOM de_PRP dados_NOM	1	58
ser_V uma_DET	17	2318
mundial_NOM de_PRP futebol_NOM	1	37
vila_NOM real_NOM	3	35
ser_V um_DET	15	2077
em_PRP espanha_NOM	4	61
direitos_NOM humanos_NOM	3	41
casa_NOM pia_NOM	2	331
amnistia_NOM internacional_NOM	1	14
ser_V o_DET	15	2999
com_PRP o_P	13	539
base_NOM de_PRP dados_NOM	2	98
and_NOM the_NOM	1	206
ter_V um_DET	9	141
lobo_NOM xavier_V	1	58
de_PRP lisboa_NOM	7	221

Palavra	**Domínios**	**Frequência**
campos_NOM de_PRP golfe_NOM	1	20
postos_NOM de_PRP trabalho_NOM	3	25
junta_NOM de_PRP freguesia_NOM	1	48
alberto_NOM costa_NOM	1	693
partido_NOM comunista_NOM	2	26
poder_NOM de_PRP compra_NOM	1	12
sobre_PRP o_P	3	13
bin_NOM laden_NOM	1	14
fernando_NOM gomes_NOM	1	21
rio_NOM de_PRP janeiro_NOM	5	68
miguel_NOM torga_NOM	2	4
conselho_NOM de_PRP ministros_NOM	1	137
vital_NOM moreira_NOM	2	22
tribunal_NOM de_PRP contas_NOM	2	34
helena_NOM roseta_NOM	3	62
faculdade_NOM de_PRP direito_NOM	1	37
lobo_NOM antunes_NOM	1	26
golpe_NOM de_PRP estado_NOM	1	12
em_PRP portugal_NOM	15	1545
carlos_NOM martins_NOM	1	33
banco_NOM de_PRP portugal_NOM	2	90
almeida_NOM santos_NOM	1	16
faculdade_NOM de_PRP letras_NOM	1	13
para_PRP o_P	15	646
qualidade_NOM de_PRP vida_NOM	4	33
ponto_NOM de_PRP vista_NOM	3	126
estado_NOM de_PRP direito_NOM	2	114
ter_V mais_ADV	3	34
ria_NOM formosa_NOM	1	23
cor_NOM de_PRP rosa_NOM	1	17
por_PRP ele_P	4	27
ser_V muito_ADV	7	163
para_PRP se_P	6	121
quartos_NOM de_PRP final_NOM	1	19
ser_V apenas_ADV	2	13
dalai_NOM lama_NOM	1	209

Palavra	**Domínios**	**Frequência**
fim_NOM de_PRP semana_NOM	5	259
pai_NOM natal_NOM	1	86
dia_NOM a_PRP dia_NOM	4	98
ser_V mais_ADV	6	73
com_PRP a_P	14	315
estar_V aqui_ADV	1	63
nossa_NOM senhora_NOM	4	797
se_P fazer_V	4	29
se_P ir_V	6	31
que_PR ser_V	13	3455
que_PR fazer_V	2	9

Bibliografia

[1] C. Banea, R. Mihalcea, and J. Wiebe. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. *In LREC*, 2008.

[2] G. Bhattacharyya and R. Johnson. *Statistical Concepts and Methods*. John Wiley & Sons, New York, 1 edition, 1977.

[3] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes, and blenders: Domain adaptation for sentiment classification. ACL 2007.

[4] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. EMNLP 2006.

[5] E. Boiy, P. Hens, K. Deschacht, and M. Moens. Automatic sentiment analysis in online text. *In Proceedings of the 11th International Conference on Electronic Publishing held in Vienna*, 2007.

[6] P. Chesley, B. Vincent, L. Xu, and R. Srihari. Using verbs and adjectives to automatically classify blog sentiment. *In Proceedings of AAAI Spring Symposium*, 2006.

[7] R. Dale, H. Moisl, and H. Somers. *Handbook of Natural Language Processing*. Marcel Dekker, New York, 1 edition, 2000.

[8] G. Dias. *Extraction Automatique d´Associations Lexicales à partir de Corpora.* PhD thesis, DI/FCT New University of Lisbon (Portugal) and LIFO University of Orléans (France), 2002.

[9] G. Dias, S. Guilloré, and J.G.P. Lopes. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *In Proceedings of 6th Conférence Annuelle du Traitement Automatique des Langues Naturelles.*, July 1999.

[10] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss analysis. *In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management, Bremen, Germany*, 2005.

[11] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, 2006a.

[12] A. Gil and G. Dias. Using masks, suffix array-based data structures and multidimensional arrays to compute positional ngram statistics from corpora. *In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics.*, July 2003.

[13] V. Hatzivassiloglou and R. McKeown. Predicting the semantic orientation of adjectives. *In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, pages 174–181, July 1997.

[14] V. Hatzivassiloglou and J.M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of 18th International Conference on Computational Linguistics*, 2000.

[15] V. Jijkoun and K.Hofmann. Generating a non-english subjectivity lexicon: Relations that matter. 2008.

[16] S. Kim and E. Hovy. Identifying and analyzing judgment opinions. *In Proceedings of the Human Language Technology Conference of the NAACL*, pages 200–207, 2006.

[17] R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. *In Proceedings of the 45th Annual Meetings of the Association of Computational Linguistics*, pages 976–983, June 2007.

[18] G. Miller. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1995.

[19] R. Mitkov. *The Oxford Hanbook of Computational Linguistics*. Oxford University Press, New York, 1 edition, 2003.

[20] C. Osgood, G. Suci, and P. Tannenbaum. The measurement of meaning. *University of Illinois Press*, 1971 [1957].

[21] S. Pais. Classification of opinionated texts by analogy. Master's thesis, Universidade da Beira Interior, August 2008.

[22] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278, 2004.

[23] B. Pang and L. Lee. *Opinion mining and sentiment analysis*, volume 2. 2008.

[24] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *In Proceedings of the 2002 conference on empirical methods in natural language processing*, 2002.

[25] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. *In Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 105–102, 2003.

[26] H. Schmid. Improvements in part-of-speech tagging with an application to german.

[27] H. Schmid. Probabilistic part-of-speech tagging using decision trees.

[28] J. Silva, G. Dias, S. Guilloré, and J.G.P. Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *In Proceedings of 9th Portuguese Conference in Artificial Intelligence.*, 1999.

[29] F. Su and K. Markert. From words to senses: a case study in subjectivity recognition. *Proceedings of Coling*, 2008.

[30] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. *In Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, 2001.

[31] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[32] P.D. Turney and M.L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical report, Technicl report ERC-1094 (NRC 44929), National Research Council of Canada, 2002.

[33] J. Wiebe. Learning subjective adjectives from corpora. *In Proceedings of the American Association for Artificial Intelligence (AAAI 2000)*, pages 735–740, 2000.

[34] J. Wiebe, R. Bruce, and T. O´Hara. Development and use of a gold standard data set for subjectivity classifications. *In Proceedings of the 37th Annual Meeting of the Association for Computacional Linguistics*, pages 246–253, June 1999.

[35] J. Wiebe, T. Wilson, and M. Bell. Identifying collocations for recognizing opinions. *In Proceedings of ACL/EACL ´01 Workshop on Collocation*, 2001.

[36] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *In Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 129–136, 2003.