FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Detecção Automática da Polaridade em Notícias Sobre Política

Manuel António da Rocha dos Santos

VERSÃO DEFINITIVA

Relatório de Dissertação
Mestrado em Multimédia

Orientador: Eugénio da Costa Oliveira

29 de Outubro de 2009

Detecção Automática da Polaridade em Notícias Sobre Política

Manuel António da Rocha dos Santos

Relatório de Dissertação

Mestrado em Multimédia

29 de Outubro de 2009

Resumo

A democratização do acesso à informação através dos meios de comunicação social, nomeadamente dos jornais on-line, faz com que a informação chegue a um público mais vasto de forma mais rápida e flexível. As notícias on-line são, cada vez mais, um meio preferencial de acesso à informação noticiosa e como todos os meios de comunicação social, influenciam a opinião pública. Assim, a análise das notícias veiculadas permite obter tendências na opinião pública. No caso de eleições políticas, os média podem influenciar os acontecimentos através da forma como expressam as notícias, dos termos que utilizam, da frequência e destaque que dão a uma notícia.

A análise automática da polaridade das notícias pode interessar tanto a candidatos e políticos em geral como a empresas de assessoria. Os interessados poderão acompanhar em tempo real a evolução da campanha eleitoral, da sua popularidade ou verificar como analisam os jornais on-line os debates e entrevistas do dia anterior.

O que propomos nesta dissertação é desenvolver um sistema automático de detecção da polaridade das entidades em notícias sobre política. Com este sistema pretende-se acompanhar a evolução da popularidade das entidades políticas e testar o grau de fiabilidade desta ferramenta na previsão de resultados eleitorais. Para isso, foram utilizadas técnicas de processamento de linguagem natural para extracção de dados e na detecção da polaridade foi implementado um sistema supervisionado de aprendizagem automática, utilizando Máquinas de Vectores de Suporte.

Para a avaliação do classificador foi realizado um esquema de validação cruzada em que foram testados 3918 exemplos equilibrados. Os resultados obtidos foram bastante satisfatório, uma vez que alcançaram uma precisão de 78% para uma abrangência de 78%.

Com a intenção de testar a utilidade deste classificador na previsão de tendências foram comparados os resultados acumulados, no período das Eleições Europeias, relativamente aos dois principais candidatos aos respectivos partidos e seus líderes. Da análise dos resultados, ao contrario da maioria das sondagens é possível prever a vitória do PSD nessas eleições.

Abstract

The democratization of access to information through the media, particularly the newspapers online, allows the information to reach a wider audience more quickly and more flexibly. The online news are increasingly becoming a preferred means of access to information and all news media, influencing public opinion. Thus, the analysis of reports enables trends in public opinion. In the case of political elections, the media can influence the events by using expressive terms and thus giving more importance to a report.

The automatic analysis of the polarity of the news may concern both the candidates and politicians in general and the business of advising. Interested parties may monitor in real time the evolution of the electoral campaign, its popularity as a review or check the papers and online debates and interviews from the previous day.

What we propose in this dissertation is to develop an automatic system for detecting the polarity of the entities in news regarding politics. This system aims to monitor the popularity of political entities and test the degree of reliability of this tool in predicting the election results. For this, we use techniques of Natural Language Processing to extract data and to detect the polarity; for this it was implemented a system of Supervised Machine Learning using Support Vectors Machines.

For the evaluation of the classifier was a cross-validation scheme was used in which 3918 balanced examples were tested. The results were quite satisfactory, since an accuracy of 78% for a recall of 78% was achieved.

With the objective of testing the function of this classifier to determine the prevision of the results accumulated in the period of elections, for the two main candidates of their parties and their leaders where compared. Unlike most surveys, with analysis of this results it is possible to predict the victory of the PSD in these elections.

Agradecimentos

Uma dissertação é fundamentalmente um trabalho individual. No entanto, existe a necessidade de um ambiente propício para que esse trabalho se vá erguendo ao longo do tempo. Este ambiente não é de forma alguma fruto de uma contribuição individual, mas sim dum conjunto de contribuições. É às pessoas que deram essa contribuição a quem quero agora agradecer:

À minha esposa e filha pela presença, compreensão e motivação. Às duas o meu muito obrigado.

Ao meu orientador, Eugénio da Costa Oliveira e ao Luís Sarmento, pela disponibilidade, pela forma como me conduziram e impulsionaram a ir sempre mais à frente. Os meus sinceros agradecimentos.

Àquelas pessoas que de alguma forma contribuíram, libertando-me para que eu pudesse dedicar todo o tempo possível a esta dissertação. A essas pessoas, o meu muito obrigado.

Manuel António Santos

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
RSS	Rich Site Summary
PERL	Practical Extraction and Report Language
SVM	Support Vector Machines

Capítulo 1

Introdução

1.1 Contexto

Com a evolução das técnicas de Processamento de Linguagem Natural e de Aprendizagem Automática, novos estudos surgem da aplicação destas tecnologias em diferentes áreas de interesse, como por exemplo a análise de sentimento em blogues sobre diversos tópicos, nomeadamente, política, tendências de bolsa de valores, opinião sobre artigos, filmes, etc. No caso do tópico política, área em que se insere esta tese, o interesse tem vindo a aumentar reflectindo-se especialmente em estudos realizados por algumas Universidades.

1.2 Motivação

Embora esteja a surgir alguns estudos sobre a classificação de texto de carácter político, esta é uma área ainda em desenvolvimento.

Com as recentes Eleições Europeias pretende-se aproveitar o elevado afluxo de notícias de carácter político, publicadas pelos meios de comunicação social on-line, para implementar e avaliar um classificador automático de notícias curtas obtidas de diferentes fontes RSS.

O recente fracasso das sondagens para as Eleições Europeias leva-nos a crer que existe a necessidade de analisar outras fontes de informação que permitam detectar tendências eleitorais. A análise da polaridade das notícias poderá vir a ser uma fonte a ter em conta se esta se revelar pertinente. Com a implementação de um sistema automático de classificação de notícias, este processo será automático e em tempo real.

1.3 Objectivos

Pretende-se implementar um classificador automático que detecte a polaridade em notícias de carácter político, a partir de fontes RSS on-line. Para a concretização e implementação deste classificador será necessário realizar as seguintes tarefas:

1. Recolha automática das notícias on-line.
2. Criação da lista de entidades políticas.
3. Identificação automática das entidades nas notícias.
4. Desenvolvimento de uma ferramenta de anotação manual das notícias, que serão utilizadas no treino e teste do classificador.
5. Anotação manual de notícias para treino e teste do classificador.
6. Geração automática dos vectores de features, a partir das notícias recolhidas, para treino e teste do classificador, como representado na Secção 3.6.
7. Treino do classificador e geração do Modelo SVM.

Depois de implementado o classificador será realizada a avaliação da Precisão em função da Abrangência, utilizando um esquema de validação cruzada, de todos os tipos de features independentemente e em conjunto, como apresentado no Capítulo 4. Para a concretização deste processo de avaliação será necessário efectuar as seguintes tarefas:

1. Selecção dos exemplos de notícias para treino e teste de forma a implementar a validação cruzada automaticamente.
2. Cálculo automático das medidas: Precisão e Abrangência.

Depois da fase de avaliação, o classificador passará à fase de produção. Nesta fase, serão automaticamente classificadas as novas notícias que diariamente são recolhidas das fontes RSS. Com os resultados da classificação serão gerados gráficos de valores acumulados, relativamente às entidades políticas.

Pretende-se utilizar os resultados das Eleições Europeias para comparar com os resultados apresentados pelos gráficos de valores acumulados e avaliar, desta forma, a robustez do sistema na previsão de resultados eleitorais.

Além destes objectivos a que nos propomos, existem outras questões relacionadas com a análise de textos de carácter político que poderão ser interessantes analisar. Um exemplo seria a identificação de desvios nos resultados dos diferentes jornais relativamente aos resultados eleitorais.

Outros temas são já estudados, nomeadamente:

- Identificação da filiação partidária do autor.
- Classificar os pontos de vista dos autores relativamente a uma facção política.
- Avaliação do grau de confiança com o qual o escritor exprime a sua opinião.
- Avaliação do grau de agradabilidade e argumentação com o qual o autor comunica.
- Identificar questões políticas de particular importância para o autor.

1.4 Estrutura da Dissertação

Para além da Introdução, esta dissertação contém mais 5 capítulos. No Capítulo 2, Revisão Bibliográfica - são apresentados artigos que, de alguma forma, estejam relacionados com o tema desta dissertação, quer pelo tema, quer por tecnologias ou técnicas utilizadas. São apresentadas as técnicas de aprendizagem automática que são utilizadas nos artigos estudados, nomeadamente Naïve Bayes e Suport Vector Machines. No final são apresentados trabalhos relacionados.

No Capítulo 3, Descrição do Sistema - é realizada a descrição do sistema em que se apresenta a sua arquitectura, define-se o corpus utilizado, são descritos os diversos tipos de features utilizados e por fim é realizada uma descrição aproximada da forma como as SVMs calculam o valor obtido da classificação.

No Capítulo 4, Avaliação - é descrito os métodos standard de avaliação do sistema. É apresentado um exemplo de classificação de uma notícias e são apresentados os resultados obtidos na classificação.

No Capítulo 5, Análise de Desempenho - são apresentados os resultados em forma de gráficos (Precisão vs Abrangência) das avaliações realizadas às features independentemente e em conjunto. São comparados os resultados das Eleições Europeias com gráficos de valores acumulados de forma testar a robustez do classificador na previsão de resultados eleitorais.

No Capítulo 6, Conclusões e Trabalho Futuro - são apresentados os resultados e as principais conclusões. É feita referência a resultados que seria interessante avaliar e são apontados novos caminhos para trabalhos futuros.

1.5 Definição de Termos

Esta lista de definição de termos, pretende facilitar a compreensão dos assuntos que serão tratados ao longo da dissertação.

webpost - registo num blogue.

anotador - pessoa que classifica e anota manualmente uma base de conhecimento, utilizada para treinar um classificador ou comparar resultados com um sistema específico.

titular - autor, responsável, detentor de uma afirmação.

tópico - tema específico escolhido na selecção do texto.

entidade - pessoa, organização, objecto ou outro do qual se expresse opinião.

bag-of-features - lista de palavras classificadas pela sua polaridade (negativo/positivo) e que poderá ter um peso associado à polaridade (o quanto é negativo e o quanto é positivo).

data mining - processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

rótulo - referência de identificação de característica pertencente a uma classe.

classe - conjunto de características que constituem uma classe.

característica (features) - forma de representação a partir da qual um classificador automático "aprende" e aplica a novos casos semelhantes, permitindo construir um sistema de aprendizagem automática.

blog - registo cronológico, frequentemente actualizado de opiniões, emoções, factos, imagens, links ou qualquer outro tipo de conteúdo que o autor ou autores queiram disponibilizar.

blogger - pessoa que publica as suas opiniões, emoções, factos, imagens, links,..., num blog ou contribui para a actualização de um blog.

léxico - dicionário dos vocábulos usados num domínio especializado (ciência, técnica)[1]. No nosso caso, um léxico é uma lista de palavras assinaladas com polaridade positiva ou negativa, que pode ter associada uma escala de polaridade e podem encontrar-se agrupadas em classes.

média social - é conteúdo informativo criado por pessoas que usam ferramentas de publicação bastante acessíveis. Destina-se a facilitar a comunicação, influenciar a interacção entre

[1] Dicionário online http://www.priberam.pt

pares e com audiências públicas. É normalmente realizado através da Internet (ex: youtube, hi5, twitter, etc.).

script - instruções escritas numa linguagem de programação que interpretadas e executadas por um computador executam acções.

stopwords - palavras que normalmente não contêm informação relevante. Como por exemplo pronomes, proposições, conjunções, etc.

bigrama - conjunto de duas palavras.

sistema supervisionado - um sistema supervisionado de aprendizagem automática aprende sendo-lhe submetido exemplos pares de entrada / saída (exemplo de teste / resultado).

Capítulo 2

Revisão Bibliográfica

2.1 Introdução

Neste capítulo é realizado o estudo de diversos artigos, quer seja pelos assuntos tratados, pelas técnicas utilizadas ou resultados obtidos. São apresentados os problemas mais abordados e estudadas as técnicas de aprendizagem automática empregues, bem como as diferentes heurísticas utilizadas. Por fim são apresentados exemplos de trabalhos que aplicam estas diferentes técnicas e tentam, de alguma forma, reflectir tendências.

2.2 Problemas Abordados

Um dos problemas mais abordados é a identificação e classificação de sentimento numa opinião. Entenda-se sentimento como a polaridade da opinião, positiva ou negativa, expressa por um titular. Kim e Hovy [KH04] descrevem uma opinião, como um conjunto constituído por Tópico, Titular, Afirmação e Sentimento no qual o Titular acredita numa afirmação sobre um Tópico e em muitos casos associa um Sentimento bom ou mau. Na abordagem a estes problemas apresentam um sistema, em que dado um tópico e um conjunto de textos, opera da seguinte forma:

1. São seleccionadas as frases que contêm o Tópico e o Titular.
2. A região do texto que diz respeito ao titular é delimitado.
3. O classificador calcula a polaridade de cada palavra.
4. Finalmente, o sistema combina o resultado com o respectivo titular, produzindo o sentimento de toda a frase.

A classificação de corpus de carácter político tem vindo a ser estudado por diversos investigadores. Estes têm experimentado diferentes abordagens ao problema e de diferentes pontos de vista. Uma das abordagens possíveis é a identificação da tendência política num corpus. Robert Malouf e Tony Mullen [MM06] pretendem demonstrar que num conjunto de posts dum grupo de discussão política, um post realizado em resposta directa a outro post tem forte tendência a representar pontos de vista opostos ao post original.

Gregory Grefenstette, Yan Qu, James G. Shanahan e David A. Evans, realizam uma abordagem ao nível da classificação de notícias sobre figuras públicas ligadas à política, através da frequência de palavras positivas e negativas [GQSE04]. Neste estudo é utilizado o browser Google News para seleccionar as notícias mais requisitadas.

Os diferentes passos desta abordagem são descritos a seguir:

- O utilizador especifica a entidade que pretende filtrar, como representado na imprensa, bem como o período de tempo envolvido.

- O sistema envia uma solicitação ao Google News que disponibiliza até 1000 notícias relativas à entidade num período de tempo específico.

- A cada notícia seleccionada é extraído o texto ao redor da entidade com recurso ao programa KWIC Keyword-in-Context (Heaps, 1978))[1]. Nesta selecção, são utilizados 120 caracteres antes e depois da entidade, como uma "janela".

- As "janelas" extraídas são ordenadas e as duplicações removidas para eliminar partes de notícias duplicadas.

- As "janelas" são recolhidas e todas as palavras, em qualquer variante morfológica do léxico, são identificadas.

- As palavras são afectadas a cada classe utilizando um léxico.

- A pontuação para a entidade é obtida pela divisão do número de referências a palavras da classe positiva pelo número de referências a palavras da classe negativa. Se houver mais referências positivas do que negativas, a pontuação será maior que um, se houverem mais referências negativas, será inferior a um.

[1] Metodologia automática de pesquisa usada para criar índices baseados no texto ou títulos de documentos. Cada palavra-chave é armazenada juntamente com parte do texto adjacente, em geral, uma palavra ou frase.

Outras abordagens, mais completas, podem considerar uma escala de pontuação de polaridade e de subjectividade. Como, por exemplo, um sistema que atribui pontuação indicando o parecer positivo ou negativo a cada entidade distinta num corpus [GSS07]. O sistema é composto por uma fase de percepção da polaridade, que associa as opiniões expressas com cada entidade e uma fase de pontuação e agregação de polaridade. A pontuação é realizada à entidade em relação a outras entidades da mesma classe. Esta pontuação é efectuada quanto à polaridade e à subjectividade.

Tendo em conta a especificidade de cada problema, podemos considerar uma abordagem que pretenda prever dia-a-dia mudanças na opinião pública sobre candidatos políticos, a partir de notícias diárias e de informação de mercado, de forma a prever a evolução de uma bolsa de apostas nas presidenciais americanas. Este sistema baseia-se no princípio de que a opinião pública é influenciada pelas notícias veiculadas pela comunicação social e pela forma como as novas notícias influenciam esta opinião. Imaginemos que no dia a seguir a um debate, a maioria dos jornais publicam notícias favoráveis a George Bush, aumentando a sua popularidade e consecutivamente o aumento do preço da cotação de George Bush. O debate é discutido por vários dias após o evento, no entanto, a popularidade de George Bush não irá continuar a subir baseado em noticias passadas. A mudança na opinião pública deve reflectir as mudanças na cobertura noticiosa diária. Em vez de construir recursos para um único dia, estes recursos podem representar diferenças entre os dois dias de cobertura noticiosa, ou seja, a novidade da cobertura. Baseados nesta informação e na evolução da bolsa de apostas Namrata Godbole, Manjunath Srinivasaiah e Steven Skiena, [LGDP08] construíram um sistema de aprendizagem automática que pretende prever o comportamento de um mercado de apostas nas presidenciais americanas.

Nas diferentes abordagens à classificação de corpus de carácter político, os blogs apresentam novos desafios à análise de texto. Mais do que somente texto, os blogs apresentam bastante informação sobre o autor, como a localização, idade, sexo, etc. Os blogs são utilizados por grupos de opinião, como por exemplo fóruns de tópico específico o que faz deles uma fonte de opinião interessante de analisar.

William W. Cohen e Frank Lin, [CL08] apresentam um sistema para classificar blogs de categoria política, desenvolvendo para o efeito um algoritmo de aprendizagem semi-supervisionado baseado no PageRank[2], a que chamaram MultiRank, para classificação de blogs políticos e classifica-os através de duas classes pré-definidas, uma de direita e outra

[2]Solução apresentada pelo Google que consiste em atribuir um valor numérico, designado por PageRank, a cada uma das páginas da Internet, de acordo com a sua "importância". Assim, as primeiras páginas a serem apresentadas ao utilizador como resultado de uma busca serão aquelas com maior valor de PageRank. Para que uma página tenha um valor elevado de PageRank, não interessa apenas que receba um grande número de ligações de outras páginas, mas também que essas páginas tenham um PageRank elevado e que esse PageRank transmitido seja partilhado com o menor número possível de outras páginas (atendendo ao facto de o PageRank transmitido por uma página ser partilhado por todos os links dessa página, receber um link de uma página que possui apenas 5 links no total pode ser mais vantajoso do que receber um link de uma página com um PageRank superior mas com, digamos, 100 links).

de esquerda. Na predição da classe do link, é explorada uma propriedade de ligação que se encontra na blogosfera política: "blogs com tendência política semelhantes tendem a ligarem-se entre si" (Adamic & Glance 2005) [AG05]. Um link de um blog de uma determinada facção política é como um link que confirma a facção. A classificação dos blogs inicializa-se a partir da propagação da tendência política, ao links identificados nos blogs, de um conjunto inicial de blogs sementes de facção conhecida.

Paula Chesley, Bruce Vincent e Li Xu, Rohini K. Srihari [CVXS06] analisaram e classificaram webposts como objectivos, subjectivos positivos ou subjectivos negativos, através da utilização de classes de verbos e adjectivos. A ideia é questionar se um post expressa subjectividade ou objectividade e no caso de expressar objectividade esta tem polaridade positiva ou negativa. Embora a classificação da subjectividade vs objectividade seja idealmente expressa numa escala contínua, simplificaram e utilizaram uma classificação binária. Neste estudo são abordadas as seguintes questões relacionadas com a análise da polaridade em blogs:

1. Como é que a utilização de classes de verbos se comportam na classificação da polaridade?

2. Como é possível utilizar recursos lexicais on-line, como o dicionário da Wikipédia, para automaticamente classificar adjectivos que expressem polaridade?

3. Qual é a melhor forma de propagar informação da polaridade do nível lexical para o nível do webpost?

4. É um blog diferente de outros tipos de fontes no que toca à expressão de polaridade, e se sim, como?

Kathleen T. Durant e Michael D. Smith [DS06] testaram duas técnicas diferentes de aprendizagem automática: Naïve Bayes e Support Vector Machines para determinar a sua aplicabilidade no domínio da classificação de blogs de categoria política. Estas técnicas de Aprendizagem Automática são métodos genéricos de classificação com sucesso em diversas áreas, como reconhecimento de face, classificação de texto, Bio-informática, análise de bases de dados, etc. Nesta experiência os autores demonstraram que um classificador Naïve Bayes consegue prever correctamente a polaridade de posts de categoria política, em média, 78,06% das vezes com um desvio padrão de 2,39. O classificador Support Vector Machines, em média, previu correctamente a categoria dos blogs 75,47% das vezes com um desvio padrão de 2.64. No entanto, este artigo refere um outro estudo que em foi possível atingir uma precisão de 82,9% com recurso a Support Vector Machines, utilizando as características escolhidas de um corpus sem tópico específico [PLV02][3].

[3] Este estudo considera o problema da classificação de documentos não por tópico, mas pela polaridade da totalidade do corpus.

2.2.1 Recursos e Suas Características Intrínsecas

A opinião das pessoas foi sempre uma informação importante na altura de tomar decisões. A Internet, através da web 2.0, é hoje um meio muito mais democrático e abrangente de expressão de opinião e informação, materializado em jornais on-line, blogs, fóruns, redes sociais, etc. A análise desta informação, torna-se bastante interessante na medida em que as pessoas divulgam a própria opinião, sem quaisquer constrangimentos. No entanto, a análise desta informação, oferece muitos desafios. A análise de uma opinião implica a identificação de um titular e de um oponente, a identificação da opinião e a sua classificação. Mas a classificação de uma opinião implica outros desafios, principalmente no caso dos blogs, uma vez que é frequente a utilização de linguagem pouco formal, a utilização de abreviaturas, imagens e "uma imagem vale mais do que mil palavras" até outros links que sequenciam o discurso.

Seguidamente são apresentadas as fontes e os tipos de dados que são utilizados nos artigos estudados:

2.2.1.1 Notícias On-line

Os jornais on-line são hoje e cada vez mais os meios mais utilizados de acesso á informação noticiosa, por isso são estes meios que mais influenciam a opinião de um publico. Os média moldam a opinião pública e por sua vez influenciam os acontecimentos. Baseado neste princípio Kevin Lerman, Ari Gilder, Mark Dredze e Fernando Pereira [LGDP08] pretendem prever, numa bolsa de apostas sobre as presidenciais americanas, se a popularidade dia-a-dia de um candidato sobe ou desce como um título numa bolsa de valores. Neste estudo são utilizados artigos noticiosos (50 artigos por dia durante 3 meses) de cobertura das eleições presidenciais obtidas da Factiva[4], um arquivo de notícias on-line mantido pelo Dow Jones, como informação externa ao mercado para prever as cotações para o dia seguinte.

Embora os jornais geralmente lutem por objectividade na comunicação, é quase impossível a utilização de palavras que não exerça algum conteúdo emocional ao descrever um acontecimento ou uma pessoa. Baseado neste princípio, um estudo de G. Grefenstette, Y. Qu, J. G. Shanahan, e D. A. Evans [GQSE04] pretende criar uma aplicação que mede a popularidade de figuras públicas a partir das opiniões vinculadas pela imprensa. Para isso, utiliza o Google News que selecciona até 1000 referências com maior ranking de consulta para artigos noticiosos relativos a uma entidade e durante um período de tempo específico.

[4] http://www.factiva.com

2.2.1.2 Blogs

O artigo [CL08] implementa um algoritmo de aprendizagem semi-supervisionado para classificação de blogs políticos e classifica-os através de uma classe pré-definida. Na predição do rótulo do link, é explorada uma propriedade de ligação que se encontra na blogosfera política: blogs com tendência política semelhantes tendem a ligarem-se entre si (Adamic & Glance 2005).

O estudo do artigo [DS06] tem como objectivo classificar blogs de categoria política como de esquerda ou direita ou sem tendência, utilizando classificadores Naïve Bayes e Support Vector Machines. Como fonte de dados para treino dos classificadores, utiliza as referência disponibilizadas no blog http://www.themoderatevoice.com, um blog sobre política que enumera e classifica mais de 250 blogs com opiniões de esquerda, direita ou opiniões moderadas. A lista foi criada pelo jornalista Joe Gandelman, que se considera como um político moderado. Foram recolhidos posts a partir de blogs de esquerda e de direita no período de Março de 2003 a Março de 2005. Os posts foram seleccionados por tópicos, identificando quais os posts que referem determinado tópico. Dos 99 blogs de esquerda e os 85 blogs de direita listados no www.themoderatevoice.com em Março de 2005, 84 blogs de esquerda e 76 de direita foram incluídos no estudo.

2.2.1.3 Notícias e Blogs

No artigo [GSS07] é comparada a polaridade, negativa / positiva, expressa numa grande quantidade de notícias e blogs, relativamente a diversas personalidades.

2.2.1.4 Notícias, Feeds e Blogs

O artigo [CVXS06] pretende classificar frases quanto à objectividade e subjectividade e à respectiva polaridade, utilizando verbos e adjectivos. Os dados foram manualmente escolhidos de posts de blogs de assuntos diversificados. Os feeds objectivos são provenientes de sites que disponibilizam conteúdos nacionais e internacionais (CNN, NPR,etc.) e jornais de notícias locais (Atlanta Journal, Constitution, Seattle Post-Intelligencer,etc.). Para a categoria de subjectivos, foram seleccionados websites tradicionais, onde somente foi tido em atenção fontes chave que permitam classificar a sua polaridade (positiva ou negativa), como notícias de jornais, cartas de editores, revistas e blogs de política.

2.2.1.5 Outros Recursos

Para testar o algoritmo que desenvolveram para detecção da polaridade, Kim e Hovy [KH04] utilizaram 100 frases de três categorias (positivas, negativas e neutras) que foram selec-

cionadas do corpus DUC 2001[5] com os seguintes tópicos: “illegal alien”,“terms limits”, “guns control” e “NAFTA”.

2.3 Técnicas de Aprendizagem Automática

Dos artigos estudados as técnicas de Aprendizagem Automática utilizadas são Naïve Bayes [DS06] [MM06] e Support Vector Machines [CVXS06] [DS06].

2.3.1 Naïve Bayes

O algoritmo Naïve Bayes determina a probabilidade de um dado elemento pertencer a uma determinada categoria C_i, para isso, calcula a probabilidade do elemento pertencer a cada uma das classes disponíveis e define aquela que tem maior probabilidade de ser a escolhida para classificação.

Como definir a qual classe um novo elemento tem mais hipóteses de ser classificado?

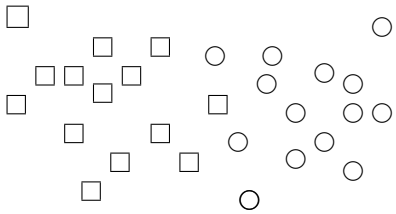

Figura 2.1: Conjuntos de características inicial

Se considerarmos que existem mais círculos do que quadrados (45 círculos e 15 quadrados), é mais provável que um novo caso se enquadre entre os círculos do que entre os quadrados. Esta probabilidade é conhecida como Probabilidade à Priori.

Definição de Probabilidade à Priori:

- P(Círculo) = Número de círculos / Número total de objectos
- P(Quadrado) = Número de quadrados / Número total de objectos

[5]Document Understanding Conference, disponibiliza corpus jornalísticos de diversas fontes classificados por um grupo de especialistas.

Considerando que existem 60 objectos, sendo 45 círculos e 15 quadrados:

- P(Círculo) = 45/60
- P(Quadrado) = 15/60

Como definimos a probabilidade de um novo objecto pertencer a uma das classe (verosimilhança).
Imaginemos que surge um novo ponto, uma cruz, entre os círculos e os quadrados e que é necessário determinar a probabilidade condicionada de pertencer a um dos grupos existentes.

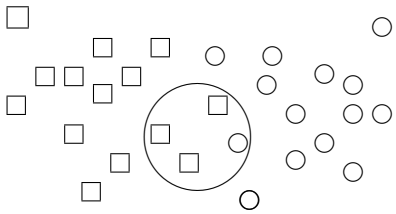

Figura 2.2: Conjuntos de características

- Supondo que os objectos estão agrupados, podemos assumir que quanto mais objectos de determinada classe estejam próximos do novo ponto, maior é a probabilidade deste ponto ser associado a essa classe.
- Para medir esta probabilidade (verosimilhança), traçamos uma circunferência em redor da cruz, isolando internamente os pontos que estão mais próximos. A partir da contagem destes pontos vizinhos, por classe a que pertencem, é possível obter a probabilidade que a cruz tem de ser círculo ou quadrado.
- P(cruz, círculo) = Número de círculos vizinhos à cruz / total de círculos.
- P(cruz, quadrado) = Número de quadrados vizinhos à cruz / total de quadrados.

- P(cruz, círculo)= 1/45
- P(cruz, quadrado)= 3/15

O teorema de Bayes funciona associando estas duas probabilidades. Na análise bayesiana, a probabilidade final é obtida da combinação de ambas as fontes de informação, formando a probabilidade à posteriori:

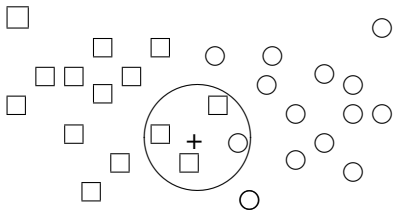

Figura 2.3: Conjuntos de características na vizinhança

- P(quadrado,cruz) = P(cruz, quadrado) * P(quadrado)
- P(quadrado,cruz) = 3/15 * 1/4 = 1/20
- P(círculo, cruz) = P(cruz, círculo) * P(círculo)
- P(círculo, cruz) = 1/45 * 3/4= 1/60

Podemos, finalmente, definir que a cruz deve ser definida como quadrado, já que a probabilidade à posteriori do quadrado foi superior.

2.3.1.1 Utilização de Naïve Bayes na Classificação de Texto

A classificação Naïve Bayes determina a probabilidade de um documento representado por um vetor de termos (características) e pesos $d_j = w_{1j}, ..., w_{|T|j}$ pertencer a uma determinada classe C_i.

$$P(c_i|\vec{d}_j) = \frac{P(c_i)P(\vec{d}_j|c_i)}{P(\vec{d}_j)}$$

$$P(\vec{d}_j|c_i) = \prod_{k=1}^{|T|} P(w_{kj}|c_i)$$

Como $P(\vec{d}_j)$ é um denominador comum, pode ser ignorado

$$P(c_i|\vec{d}_j) = \prod_{k=1}^{|T|} P(w_{kj}|c_j)P(c_i)$$

Definição de $P(w_k|c_j)$

$P(w_k|c_j) = \left(\frac{n_k+1}{n+|Lexico|}\right)$, onde:

- $P(w_k|c_j)$ - representa a probabilidade da verosimilhança da evidência do termo w_k dada a hipótese da classe c_j.
- n_k - representa a quantidade de vezes que o termo w_k aparece no conjunto de treino designado na classe c_j.
- n - representa o total de termos que fazem parte do conjunto de treino da classe c_j.
- Léxico - representa o total de termos encontrados nos dados de treino de todas as classes.

Nos artigos [DS06] e [MM06] é testado um classificador Naïve Bayes na classificação, em duas classes uma de esquerda e outra de direita, de blogs de categoria política. Um classificador Naïve Bayes é um classificador probabilístico baseado em modelos que incorporam forte independência entre as características. Nos casos aqui colocados o classificador Naïve Bayes atribui a um determinado post d uma classe c *.

$c^* = Argmax_c P(c|d)$;
$c \in \{features\ de\ direita,\ features\ de\ esquerda\}$.

Um documento de comprimento n representado como um vector dimensional m, onde f_i é a dimensão no vector e m o número de características. Derivando o classificador Naïve Bayes pela observação da primeira regra de Bayes.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

$P(d)$ não desempenha qualquer papel na atribuição $c*$. Para estimar o termo $P(d|c)$, Naïve Bayes decompõe assumindo que todas as *fi*'s são condicionalmente independentes dada a classe d:

$$P_NB(c|d) = \frac{P(c)(\prod_{i=1}^{m})P(f_i|c)^{n_i(d)}}{P(d)}$$

O classificador Naïve Bayes assume a independência dos atributos na classificação de texto em função do nível da palavra. Quando o número de recursos é grande, a independência permite que os parâmetros de cada característica sejam aprendidos separadamente, simplificando muito o processo de aprendizagem.
É utilizado um conjunto de características de palavras para representar os webposts.

$$\vec{d} = (f_1(d), f_2(d), ..., f_m(d))$$

Se $\{f_1, ..., f_m\}$, for um conjunto pré-definido de m características que podem aparecer num post. Então $f_i(d)$ é igual a 1 se o recurso f_i aparecer no post d, e é igual a 0 se a característica f_i não aparece no post d. Em seguida, cada post d é representado pelo vector post $\vec{d}$.

2.3.2 Support Vector Machines

As SVMs são um método de aprendizagem introduzido por V. Vapnik [Vap95] [CV95][BGV92]. Existem diversos tipos de SVMs mas, nesta abordagem, estudaremos apenas as SVMs lineares.

As SVMs lineares com margens rígidas definem fronteiras lineares a partir de dados linearmente separáveis. Seja T um conjunto de treino com n características, $x_i \in X$ e y_i as classes a que pertencem com $y_i \in Y$, em que X constitui o espaço dos dados e $Y = \{-1, +1\}$. T é linearmente separável se for possível separar os dados das classes +1 e -1 por um hiperplano.

Os classificadores que separam os dados por meio de um hiperplano são denominados lineares. A equação de um hiperplano é apresentada na Equação 2.1, em que $w.x$ é o produto escalar entre os vectores w e x, $w \in X$ é o vector normal ao hiperplano descrito e $\frac{b}{\|w\|}$ corresponde à distância do hiperplano em relação à origem, com $b \in R$.

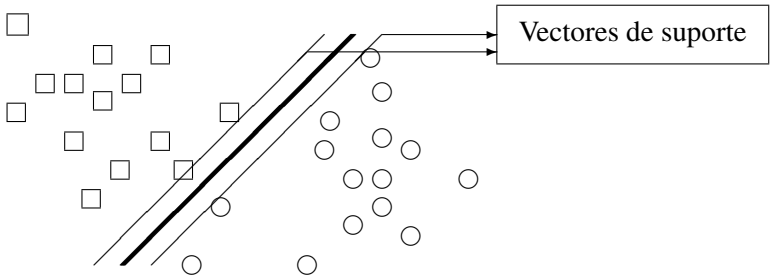

Figura 2.4: Conjuntos de treino linearmente separáveis

$$f(x) = w \cdot x + b = 0 \tag{2.1}$$

A Equação 2.1 divide o espaço dos dados X em duas regiões: $w \cdot x + b > 0$ e $w \cdot x + b < 0$.

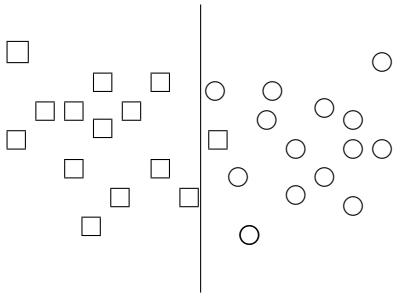

Figura 2.5: Conjuntos de treino linearmente inseparáveis

Uma equação em função do sinal $g(x) = sgn(f(x))$ pode então ser aplicada na obtenção das classificações, conforme é ilustrado na Equação 2.2.

$$g(x) = sgn(f(x)) = \begin{cases} +1 \text{ se w . x + b} < 0 \\ -1 \text{ se w . x + b} > 0 \end{cases} \tag{2.2}$$

A partir de $f(x)$, é possível obter um número infinito de hiperplanos equivalentes, pela multiplicação de w e b por uma mesma constante. Define-se o hiperplano canónico em relação ao conjunto T como aquele em que w e b são apresentados de forma a que os exemplos mais próximos ao hiperplano $w \cdot x + b = 0$ satisfaçam a Equação 2.3.

$$|w \cdot x_i + b| = 1 \tag{2.3}$$

Esta equação implica as inequações 2.4.

$$\begin{cases} w \cdot x_i + b \geq +1 \text{ se } y_i = +1 \\ w \cdot x_i + b \leq -1 \text{ se } y_i = -1 \end{cases} \tag{2.4}$$

Se x_1 é um ponto no hiperplano $H_1 : w \cdot x + b = +1$ e x_2 é um ponto no hiperplano $H_2 : w \cdot x + b = -1$, conforme ilustrado na Figura 2.6. Projectando $x_1 - x_2$ na direcção de w, perpendicular ao hiperplano separador $w \cdot x + b = 0$, é possível obter a distância entre

os hiperplanos H_1 e H_2. Esta projecção é apresentada na Equação 2.5.

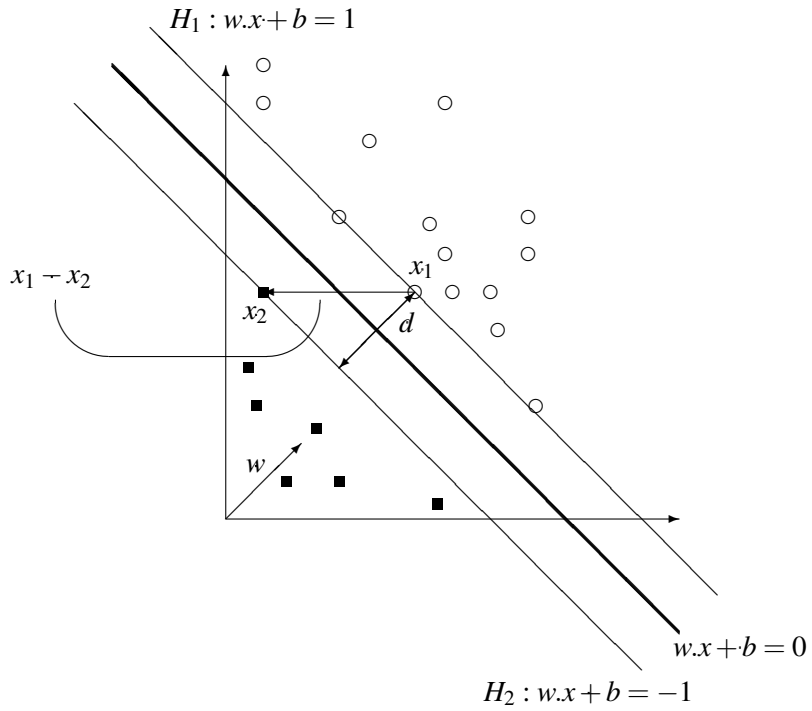

Figura 2.6: Cálculo da distância d entre os Hiperplanos H_1 e H_2

$$(x_1 - x_2)\left(\frac{w}{\| w \|} \cdot \frac{(x_1 - x_2)}{\| x_1 - x_2 \|}\right) \tag{2.5}$$

Obtem-se $w \cdot x_1 + b = +1$ e $w \cdot x_2 + b = -1$. A partir da diferença entre estas equações obtem-se $w \cdot (x_1 - x_2) = 2$. Substituindo este resultado na Equação 2.5 resulta:

$$\frac{2(x_1 - x_2)}{\| w \| \| x_1 - x_2 \|} \tag{2.6}$$

Como se pretende obter o comprimento do vector projectado, através da norma da Equação 2.6, obtem-se:

$$\frac{2}{\| w \|} \tag{2.7}$$

Esta é a distância d, ilustrada na Figura 2.6, entre os hiperplanos H_1 e H_2 , paralelos ao hiperplano separador. Como w e b foram calculados de forma a não existirem exemplos entre H_1 e H_2, w é a distância mínima entre o hiperplano separador e os dados de treino. Esta distância é definida como a margem geométrica do classificador linear. A partir destas considerações, verifica-se que a maximização da margem de separação dos dados em relação a $w \cdot x + b = 0$, pode ser obtida pela minimização de w. Dessa forma, recorre-se ao seguinte problema de optimização:

$$Minimizar_{w,b} \frac{1}{2} \| w \|^2 \tag{2.8}$$

$$\text{Com as seguintes restrições: } y_i(w.x_i + b) - 1 \geq 0, \forall i = 1, ..., n \tag{2.9}$$

As restrições são impostas de maneira a assegurar que não haja dados de treino entre as margens de separação das classes.

Os artigos [CVXS06] e [DS06] utilizam SVMs na classificação de blogs em que se identifica um hiperplano que separa duas classes de dados. O hiperplano escolhido cria a maior margem ou separação entre as duas classes, pelo que é considerado um classificador de grande margem. Assumindo que temos n blogs a serem classificados em categorias, o conjunto de webposts é representada da seguinte forma:

$$\{(x_1, c_1), ..., (x_n, c_n)\}$$

onde os x_i representam a característica do post e os c_i a categoria do post, tanto da esquerda como da direita. A divisão do hiperplano das duas classes é definida como $w.x + b = 0$. O vector de apoio para uma categoria é definido com o $w.x - b \geq 1$ e para a outra categoria $w.x - b \leq -1$. Estas desigualdades podem ser reescritas da seguinte forma:

$$c_i(w.x_i - b) \geq 1 \text{ para } 1 < i < n$$

uma vez que os c_i representam a categoria dos webposts. É um problema de optimização quadrática ao minimizar o comprimento de w dado o constrangimento acima referido. Isto irá identificar a maior margem esquerda e direita entre as opiniões. Nem todos os dados que estão classificados, são utilizados para identificar o hiperplano. Apenas são considerados os pontos mais próximos da margem das duas classes. Estes pontos são usados para determinar os vectores de apoio relativamente ao hiperplano.

Na representação dos vectores de características que representam os documentos (webposts, blogs, notícias, reportagens) a classificar pode ser utilizado:

- um típico *bag-of-features*, constituído por duas classes uma de características positivas e outra de características negativas, que podem surgir no documento que se pretende classificar. Se uma característica f_i, que faça parte do *bag-of-features*, surgir no documento d então $f_i(d)$ é igual a 1 se a característica não surgir no documento então $f_i(d)$ é igual a 0. Depois o documento d é representado por um vector de características $\vec{d} = f_1(d), f_2(d), ...f_m(d)$ [DS06].
- classes de verbos, adjectivos, nomes ou ainda combinação destas classes [CVXS06].
- as características que compõem as classes podem ser obtidas a partir de recursos web, de auto descrições partidárias dos Titulares e de discursos políticos publicados on-line [MM06].
- utilizar ferramentas especializadas, como por exemplo o InfoXtract, um analisador de texto automático que agrupa verbos de acordo com a polaridade da classe ou utilizar o dicionário da Wikipedia[6] para determinar a polaridade das características encontradas ao longo dos webposts [CVXS06].

Na Tabela 2.1 é apresentado um quadro resumo das diferentes técnicas de aprendizagem automática utilizadas nos artigos estudados.

Tabela 2.1: Técnicas de aprendizagem automática utilizados nos artigos estudados

Técnicas de Aprendizagem Automáticas	Artigos
SVMs	[CVXS06] [DS06]
NB	[DS06] [MM06]

2.4 Heurísticas Utilizadas na Classificação

Véronis[7], num estudo que realizou às eleições presidenciais francesas de 2007, relaciona o número de citações aos candidatos, efectuadas pela imprensa on-line, com o

[6] http://en.wiktionary.org/wiki/

[7] Jean Véronis é professor de linguística e ciência da computação na Universidade de Provence, onde dirige o Centro de Informática para Humanidades e Ciências Sociais. É também consultor em inúmeras empresas, membro de diversos

sucesso dos candidatos. A taxa de citações aos candidatos admitidos permitiu prever o resultado final com maior precisão do que os dados das sondagens.

Uma técnica utilizada na classificação de documentos baseia-se na relação da frequência de palavras positivas e negativas, representadas num léxico previamente definido. Se um documento é constituído por mais palavras positivas do que negativas este documento será classificado como positivo se, pelo contrário, for constituído por mais palavras negativas, então será classificado como negativo. Poderá eventualmente ser considerado neutro se existirem tantas palavras positivas como negativas [GQSE04] e [LGDP08].

Exemplo de um léxico utilizado na classificação de palavras:

Palavra	Classe	Sinal
admonish	warning	-
admonishment	warning	-
admonition	warning	-
adorable	attraction	+
adoration	love	+
doration	superiority	+
adore	love	+
adulterer	immorality	-
adultery	immorality	-
advantage	advantage	+
advantage	superiority	+

Uma abordagem mais completa atribui um peso à polaridade das palavras que constituem o texto. Nesta abordagem, são testados diferentes modelos baseados na média de sinónimos de palavras pertencentes às classes de um léxico inicialmente definido. Estes sinónimos são obtidos com a ajuda do WordNet [8].

$$argmaxP(c|w) = argmaxP(c)P(w|c)$$

$$argmaxP(c|w) = argmaxP(c)\frac{\sum_{i=1}^{n} count(syn_i,c)}{count(c)}$$

Para calcular a probabilidade $P(c|w)$ da palavra w dada a classe c, contam-se as ocorrências

comités de peritos internacionais e presidente da Associação para o Processamento de Linguagem Natural (Atala). http://sites.univ-provence.fr/veronis/

[8] A WordNet é uma base de dados de informação linguística susceptível de ser utilizada em várias áreas, tais como tradução automática, sistemas de pesquisa e de extracção de informação, sistemas periciais, aplicações para o ensino do Português, entre outras. Pode ser também utilizada como dicionário, em consultas sobre o significado das palavras e a respectiva equivalência em Inglês.

de sinónimos da palavra *w*. É também incorporado um peso (positivo / negativo) nas palavras do léxico de forma encontrar uma medida de polaridade.

$abysmal : NEGATIVE$
$|+ : 0.3811||- : 0.6188|$

$adequated : POSITIVE$
$|+ : 0.9999||- : 0.0484e-11|$

$afraid : NEGATIVE$
$|+ : 0.0212e-04||- : 0.9999|$

$P(c|s) = \frac{1}{n(c)} \sum_{i=0}^{n} p(c|w_i)$, se $argmax_j\ p(c_j|w_i) = c$

Onde $p(c|w_i).n(c)$ é o número de palavras em que a classe é c. Se contém mais palavras e com maior peso positivo de que negativas o sentimento será positivo [KH04].

$$P(c|s) = 10_{n(c)-1} x \prod_{i=1}^{n} p(c|w_i), \text{ se } argmax_j\ p(c_j|w_i) = c$$

No artigo [CL08] é proposto um algoritmo baseado no peso atribuído pelo PageRank (Page et al. 1998) que é utilizado para determinar a "importância" ou a "reputação" de um web site. A partir deste conceito é apresentada uma versão adaptada do PageRank, na qual cada site e todas as inter-ligações dos sites, está associada a uma comunidade diferente e a pontuação da "importância" propaga-se apenas no seio de uma comunidade. No contexto dos blogs políticos, a cada blog e a cada hiperlink será atribuído a uma facção particular, por exemplo, liberal ou conservador.

A seguir é descrito o método utilizado na atribuição de blogs às facções, dado um pequeno conjunto de sementes. Para avaliar a medida de reputação de uma facção específica, define-se MultiRank da seguinte forma:

$$r_f = (1-d)u + dW_f r_f$$

r_f ranking da facção em que W_{fij} é W_{ij} se o link de i para j se encontra em E_f (página inicial de uma facção) senão é zero; e u é um vector uniforme personalizado onde $u_i = 1/|V|$ e d é um factor de amortecimento constante. Nesta equação, r_f pode ser visto como a probabilidade de uma busca aleatória em G se somente seguirmos as arestas perten-

centes à facção f. Em que $G = (V,E)$, V é o conjunto inicial de blogs e E é a facção correspondente. No contexto de uma rede de blogs de política, podemos ver isto como a probabilidade de um utilizador clicar num blog liberal ou conservador aleatoriamente. Neste método é utilizada uma fase exploratória em que simplesmente classifica os links que não se encontram classificados na mesma categoria da pagina que lhe deu origem.

No artigo [LGDP08] é utilizado um típico *bag-of-features* na classificação das notícias e aplicada uma regra de regressão para reflectir a influência de um novo acontecimento na opinião pública. Num dia depois de um debate, a maioria dos jornais pode publicar que George Bush saiu vencedor do debate, originando a valorização da cotação Bush numa bolsa de apostas nas presidenciais americanas. No entanto, embora a discussão sobre o debate se possa prolongar por vários dias após o evento, os apostadores não irão continuar a apostar em Bush baseado-se em notícias passadas. Como se pretende medir a evolução da opinião pública sobre candidatos políticos de forma a integrar um sistema de apostas nas eleições presidenciais americanas, que permita prever a evolução da popularidade dos candidatos dia-a-dia, com base em notícias on-line. Para prever a alteração na opinião pública, esta deve reflectir as mudanças na cobertura noticiosa diária. Em vez de construir recursos para um único dia, pode-se representar diferenças entre os dois dias de cobertura noticiosa, ou seja a novidade da cobertura noticiosa. Dadas as contagens do recurso i no dia t como c_i^t, onde a característica i pode ser o "escândalo", e no conjunto de funcionalidades no dia t C^t, a fracção de novos focos para cada recurso $f_i^t = \frac{c_i^t}{|C^t|}$. Os novos focos mudam $(\triangle)$ para recurso i no dia t definido como

$$f_i^t = \log\left(\frac{f_i^t}{\frac{1}{3}(f_i^{t-1} + f_i^{t-2} + f_i^{t-3})}\right)$$

onde o numerador é o foco de noticias para um recurso i (hoje) e o dominador a média de focos sobre os três dias anteriores. O valor resultante capta a mudança de atenção do dia t, onde um valor maior que 0 significa maior atenção e um valor inferior a 0 diminui a atenção.

O artigo [GSS07] utiliza um sistema de classificação de todo o corpus através da contagem de palavras positivas e negativas que fazem parte de um léxico especialmente concebido para o efeito. Às palavras precedidas por uma negação é revertida a polaridade. A polaridade é incrementada ou decrementada quando a palavra é precedida por um modificador.

Ex: "not good = -1; good = 1; very good = +2".

Na Tabela 2.2 é apresentado um quadro resumo das diferentes heurísticas utilizadas nos artigos estudados.

Tabela 2.2: Diferentes Heurísticas Utilizadas

Heurísticas	Artigos
Frequência de citações às entidades	Véronis 2007
Relação da frequência de palavras positivas e negativas representadas num léxico previamente definido.	[GQSE04] [LGDP08] [GSS07]
Algoritmo baseado no peso atribuido pelo PageRank, é utilizado para determinar a importância ou a reputação de um web site.	[CL08]
Aplica uma regra de regressão para reflectir a influência de um acontecimento na opinião pública e utiliza um típico bag-of-features na classificação de documentos	[LGDP08]
Às palavras encontradas no texto precedidas por uma negação, que façam parte do léxico é revertida a sua polaridade.	[GSS07]

2.5 Técnicas de Avaliação

As técnicas de avaliação visam aferir a precisão de métodos e técnicas utilizadas na classificação, para numa fase posterior analisar os resultados. Na avaliação dos resultados dos métodos de classificação, dependendo da especificidade da técnica de classificação, do tipo de corpus e do objectivo a que a classificação se propõe, esta é realizada comparando os resultados da aplicação da técnica de classificação com os resultados obtidos por meio de anotação manual, ou comparando com os resultados da classificação em documentos que à partida é sabido o resultado.

No caso da avaliação comparando o resultado da classificação automática com os resultados da avaliação manual, o procedimento mais utilizado é a classificação de palavras pelos anotadores. Normalmente baseia-se em assinalar cada palavra em duas ou três categorias positivas, negativas ou neutras. Das palavras comuns aos anotadores é seleccionado aleatoriamente um conjunto de palavras, adjectivos, verbos ou ambos. Normalmente são comparados os resultados da classificação com alternâncias das categorias de palavras (por ex: comparar os resultados considerando as três categorias de palavras com os resultados considerando apenas duas categorias) [KH04].

Quando a classificação é realizada ao nível do documento, cabe aos anotadores classificar todo o documento. Esta classificação poderá ser uma classificação de polaridade (positivo / negativo), ou então quanto à objectividade (objectivo / subjectivo) e ainda

uma classificação quanto à polaridade da objectividade (objectivo-positivo / objectivo-negativo) [CVXS06].

Para medir a concordância entre anotadores é utilizada a estatística de Kappa (K). A Estatística K é uma medida de concordância usada em escalas nominais, que fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando assim o quão legítimas são as interpretações.

1. Primeiro calcula-se um índice que represente a percentagem de concordância esperada pelo acaso.

2. Em segundo lugar, calcula-se a concordância observada.

3. Obtidos estes dois índices, a estatística K será calculada através da divisão da diferença entre a concordância observada e a concordância esperada pelo acaso, pela diferença entre a concordância absoluta e a concordância esperada pelo acaso (a maior diferença possível entre concordância observada e esperada).

Desta forma, os valores da Estatística K variam entre 0 a 1 sendo que 0 representa não haver concordância além do puro acaso, e 1 representa a concordância perfeita. As directrizes para a interpretação (sempre subjectiva) de K são dadas na Tabela 2.3.

Tabela 2.3: Resultado da aplicação do método

Valores de Kappa	Concordância
0	Pobre
0 – 0,20	Ligeira
0,21 – 0,40	Considerável
0,41 – 0,60	Moderada
0,61 – 0,80	Substancial
0,81 – 1	Excelente

Uma outra técnica utilizada na avaliação dos métodos de classificação de texto é a aplicação do classificador em textos previamente seleccionados. Estes textos, poderão ser favoráveis ou desfavoráveis a uma determinada entidade e de tópico específico ou não. Os textos para teste, normalmente, são extraídos de sites que se sabe à partida serem favoráveis ou desfavoráveis a determinada entidade. É lógico que se extraia um texto favorável ao presidente dos Estados Unidos da América do site oficial da Casa Branca [GQSE04]. Esta técnica é também aplicada para testar a eficácia do método de classificação de texto na previsão da filiação política utilizando classificadores Support Vector Machines e Naïve Bayes [MM06]. Na escolha dos textos é também possível utilizar ferramentas que garantam a escolha dos textos com maior PageRank e ricos em citações [CL08].

O PageRank é uma solução apresentada pelo Google que consiste em atribuir um valor numérico, designado por PageRank, a cada uma das páginas da Internet, de acordo

com a sua "importância". Assim, as primeiras páginas que são apresentadas ao utilizador como resultado de uma pesquisa serão aquelas com maior valor de PageRank. Para que uma página tenha um valor elevado de PageRank, não interessa apenas que receba um grande número de ligações de outras páginas, mas também que essas páginas tenham um PageRank elevado e que esse PageRank transmitido seja partilhado com o menor número possível de outras páginas (atendendo ao facto de o PageRank transmitido por uma página ser partilhado por todos os links dessa página. Receber um link de uma página que possui apenas 5 links no total pode ser mais vantajoso do que receber um link de uma página com um PageRank superior mas com, digamos, 100 links). A fórmula original de cálculo do PageRank, desenvolvida pelos próprios fundadores do Google (Larry Page e Sergey Brin), é a seguinte:

$$PR(P_i) = (1-p) + p\left(\frac{PR(P_{j_1})}{C(P_{j_1})} + \frac{PR(P_{j_2})}{C(P_{j_2})} + ... + \frac{PR(P_{j_k})}{C(P_{j_k})}\right) \tag{2.10}$$

onde p é um parâmetro compreendido entre 0 e 1 (habitualmente p = 0,85). $PR(P_i)$ designa o valor do PageRank da página P_i e j_1, j_2, ..., j_k são os índices das páginas que possuem um link para a página P_i. $PR(P_{j_1})$, $PR(P_{j_2})$, ..., $PR(P_{j_k})$ designa o valor do PageRank dessas páginas. $C(P_{j_1})$, $C(P_{j_2})$,..., $C(P_{j_k})$ designa o número de links dessas páginas. Ou seja, ao valor mínimo do PageRank atribuído a cada página dado por $1-p$, juntamos parcelas que resultam do modo como a página P_i está relacionada na rede da Internet.

Um método utilizado para avaliar classificadores é o *K-fold Cross Validation*, que consiste em dividir o conjunto inicial de dados C em K conjuntos $K_1, K_2, ..., K_n$ com aproximadamente a mesma dimensão e a mesma classe de distribuição. A cada um destes conjuntos é aplicada uma abordagem *train and test*, em que o classificador é treinado em cada conjunto $C - K_i$ e testado no conjunto K_i. O resultado final é obtido a partir da média dos resultados individuais obtidos [DS06] [MM06] [CL08]. É habitual a utilização das médias obtidas no cálculo da precisão em função da abrangência. A precisão define-se como o valor da fracção entre o número de exemplos que o classificador acertou pelo número de exemplos testados. A abrangência define-se como a fracção entre o número de exemplos que o classificador acerto pelo número total de exemplos.

Em casos em que se pretende prever diariamente alterações na evolução da opinião (por exemplo prever a evolução da cotação da popularidade de um candidato político para o dia seguinte (como, por exemplo, uma bolsa de apostas). Em casos como este a avaliação pode ser aferida com os resultados reais do dia seguinte [LGDP08].

2.6 Quadros Resumo e Área de Intervenção da Tese

Nesta secção são apresentados quadros resumo do enquadramento dos artigos estudados, relativamente aos recursos utilizados, às técnicas de aprendizagem automática e técnicas de avaliação de resultados. Nestes quadros são assinaladas as áreas de acção desta tese.

Tabela 2.4: Quadro resumo dos recursos utilizados

Recursos	Artigos
Blogs	[CL08] [DS06]
Notícias on-line	[LGDP08] [GQSE04]
Feeds RSS	*
Notícias on-line e Blogs	[GSS07]
Notícias on-line, Feeds RSS e Blogs	[CVXS06]
Outros Recursos	[KH04]
*** - Áreas de acção da tese.**	

Tabela 2.5: Quadro resumo das técnicas de aprendizagem automática

Técnicas de Apredizagem Automática	Artigos
Naïve Bayes	[DS06] [MM06]
Support Vector Machines	[CVXS06][DS06] *
*** - Áreas de acção da tese.**	

Tabela 2.6: Quadro resumo das técnicas de avaliação

Técnicas de Avaliação	Artigos
k-fold cross validation	[DS06] [MM06] [CL08] *
Comparação dos resultados considerando números de categorias diferentes	[KH04]
Anotação manual (anotadores classificação documentos que servirão para teste)	[CVXS06]
Utilização de documentos que é sabido à partida a tendência (ex: filiação política)	[MM06]
Avaliação realizada através dos valores reais posteriormente disponibilizados	[LGDP08]
*** - Áreas de acção da tese.**	

2.7 Exemplos de Trabalhos

Alguns exemplos de trabalhos que utilizam técnicas de processamento de linguagem natural e aprendizagem automática, serão agora apresentados.

2.7.1 MoodViews

O MoodViews é um conjunto de ferramentas que monitorizam o estado de espírito em textos disponibilizados pelo LiveJournal. Actualmente, o MoodViews é constituído por três componentes, cada componente oferece uma visão diferente do estado de espírito mundial:

- Moodgrapher - estima o nível de estado de espírito mundial.
- Moodteller - prevê o nível de estado de espírito mundial.
- Moodsignals - auxilia na compreensão das razões subjacentes à evolução do estado de espírito.

2.7.1.1 MoodViews

MoodViews foi desenvolvido pela Moodteam, um grupo de investigadores que se dedicam à pesquisa de informação, nomeadamente informação web. Do ponto de vista do acesso à informação, os blogs oferecem muitas opções além das tradicionais pesquisas, tais como: detecção de tendências, análise sobre tópicos específicos, links, criação de feeds, etc. A maioria dos blogs permitem entradas através de tags específicas e pessoais. Os utilizadores do LiveJournal, tem a opção de seleccionar um estado de espírito de uma lista predefinida de 132 opções comuns, como por exemplo: divertido ou zangado ou entrada livre de texto. Uma grande percentagem de utilizadores do LiveJournal utilizam esta opção, etiquetando os posts com informação do seu estado de espírito. Isto resulta num conjunto de centenas de milhares de posts etiquetados com a informação do estado de espírito, provenientes de todo o mundo. O tipo de informação que pretendem obter é facilmente perceptível através das perguntas seguintes:

- Como evolui o estado de espírito?
- Como se relaciona o estado de espírito?
- Terão os acontecimentos mundiais impacto no estado de espírito?
- O estado de espírito global pode ser obtido através dos acontecimentos mundiais?

2.7.1.2 Moodgrapher

O Moodgrapher baseia-se no MoodViews, que é o componente de base do sistema. O Moodgrapher agrega numa interface os níveis de estado de espírito ao longo do tempo. Através desta interface reflecte padrões de estado de espírito, baseando-se nos acontecimentos com implicações mundiais.
Moodgrapher foi lançado em Junho de 2005.
Conceito e design: Gilad Mishne.
Aplicação: Gilad Mishne e Krisztian Balog.

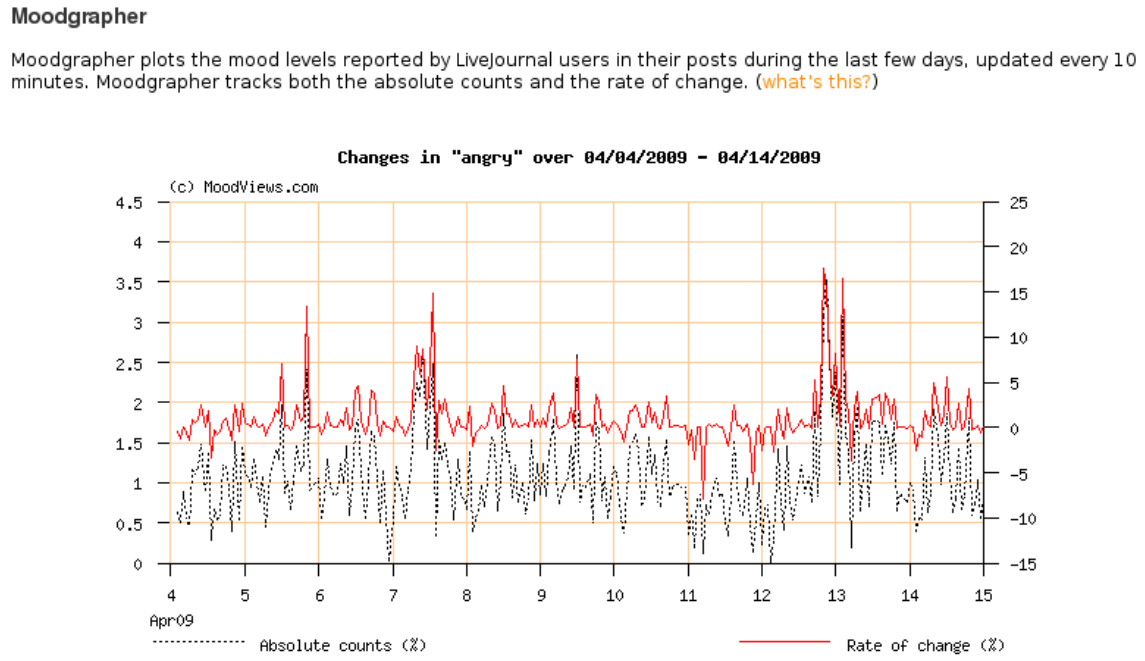

Figura 2.7: Interface Moodgrapher

Figura 2.8: Interface Moodgrapher - Parâmetros de selecção

2.7.1.3 Moodteller

O Moodteller utiliza técnicas de processamento de linguagem natural e aprendizagem automática para estimar os níveis de estado de espírito nos textos dos posts colocados no LiveJournal, sem utilizar as tags fornecidas pelos utilizadores. A estimativa é então apresentada e são comparados os valores reais com base nas etiquetas fornecidas pelos utilizadores. Por fim é apresentada a informação da precisão. O Moodteller foi lançado em Setembro de 2005.
Conceito, design e implementação: Gilad Mishne.

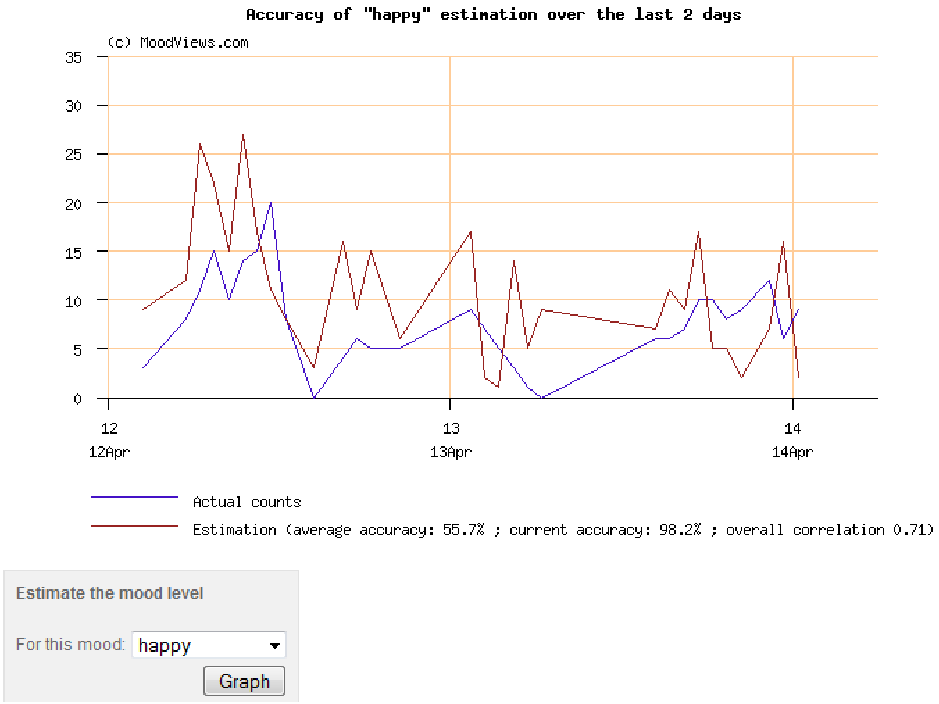

Figura 2.9: Interface do Moodteller

2.7.1.4 Moodsignals

Moodsignals detecta, num determinado intervalo de tempo, palavras e frases que são associados a um estado de espírito. O Moodsignals identifica picos de estado de espírito e tenta explicar os picos encontrados através da análise das notícias.
Moodsignals foi lançado em Março de 2006.
Conceito: Gilad Mishne.
Design: Gilad Mishne e Krisztian Balog.
Implementação: Krisztian Balog Breyten e Ernsting.

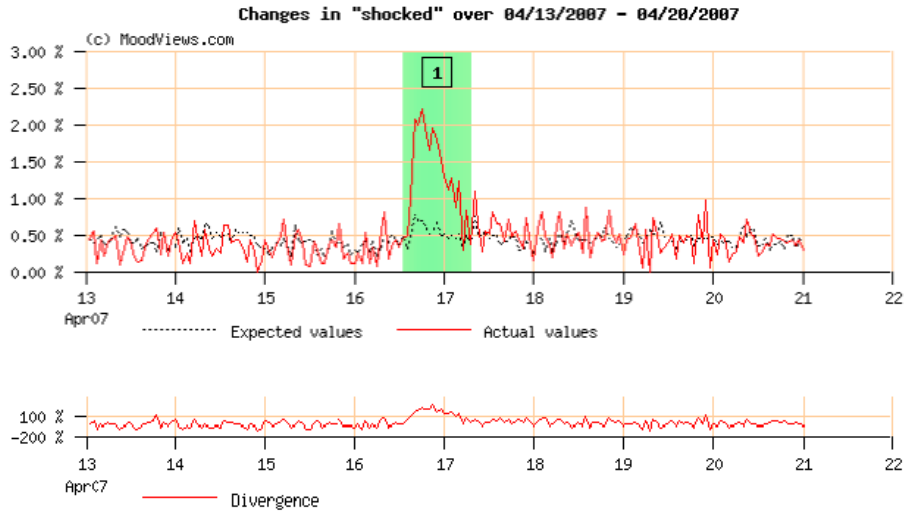

Figura 2.10: Interface do Moodsignals

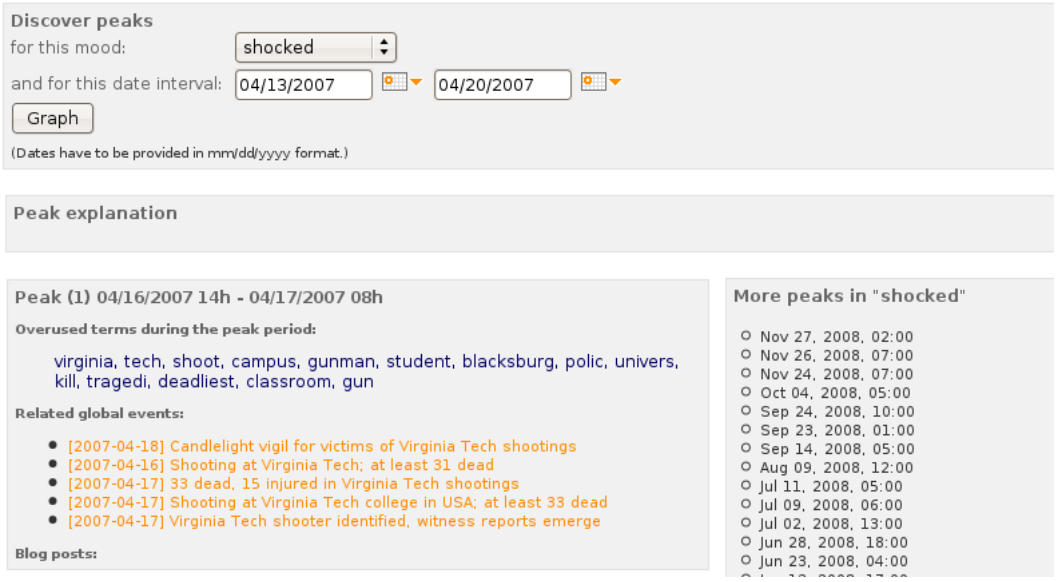

Figura 2.11: Interface do Moodsignals - Parâmetros de selecção e noticias que justificam o pico

2.7.2 Google In Quotes

O In Quotes permite pesquisar citações em notícias a partir do Google News. Estas citações são um recurso valioso para a compreensão da opinião que as pessoas que citam têm sobre os diversos assuntos. Grande parte da informação publicada, sobre estas pessoas, é baseada na interpretação de um jornalista. As citações directas, por outro lado, são unidades de informações concretas que descrevem como as pessoas se apresentam por si próprias. O Google News reúne as citações on-line, a partir de notícias, e classifica-as em grupos pesquisáveis com base nas citações. À semelhança da selecção de conteúdos a partir do Google News, a selecção de citações e os seus responsáveis são realizadas automaticamente, não garantindo a integridade ou exactidão da informação que é apresentada. As datas apresentadas são as datas em que os artigos são adicionados ao Google News de onde são extraídas as citações. A funcionalidade de comparação de citações permite comparar citações de pessoas diferentes em notícias sobre um determinado tópico. Actualmente permite escolher e comparar citações de candidatos políticos e outras figuras políticas[9].

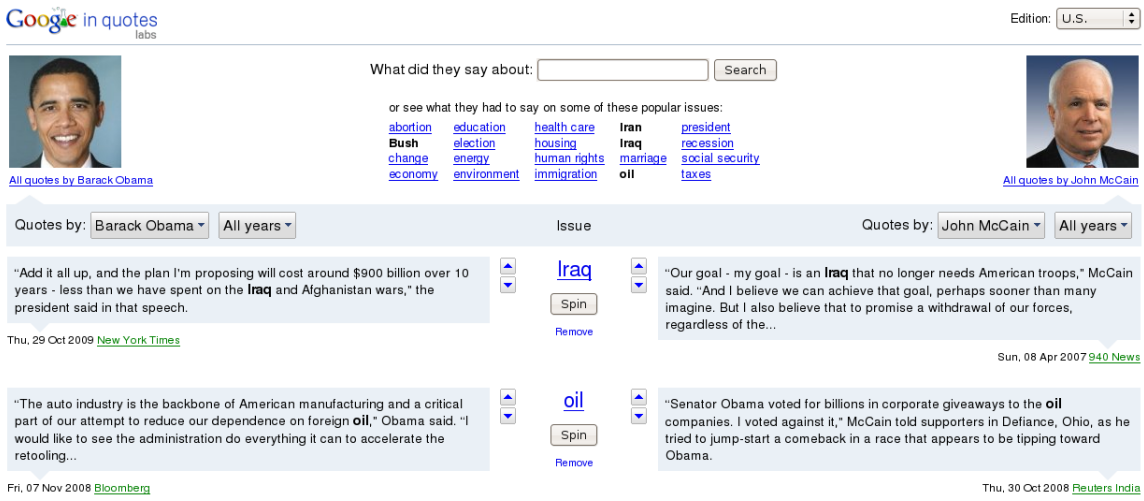

Figura 2.12: Interface do Google In Quotes

[9] http://labs.google.com/inquotes/

2.7.3 10 x 10

O 10x10 (dez por dez) pretende analisar, interactivamente, palavras e imagens que "definam" o tempo através da montagem instantânea de imagens, palavras e frases obtidas a partir de fontes de todo o mundo. A cada hora, o 10x10 recolhe as 100 palavras e imagens com maior "importância", a partir das diferentes fontes noticiosas e apresenta-as como se de uma imagem única se tratasse. Ao longo do tempo, forma um mosaico contínuo de acontecimentos constituído por imagens e palavras. O 10x10 obtém e extrai de feeds RSS as principais notícias internacionais e realiza a análise textual do resumo da notícia. As 100 primeiras palavras são recolhidas, juntamente com 100 imagens correspondentes. As notícias e as imagens são obtidas de fontes de noticiosas como a Reuters World News a BBC World Edition ou o New York Times International News. No fim de cada dia, mês e ano, o 10x10 olha para trás através dos arquivos para seleccionar as 100 palavras que melhor representem um determinado período de tempo. Uma constante evolução do mundo é formada e guardada, com base em acontecimentos mundiais importantes, sem qualquer intervenção humana[10].

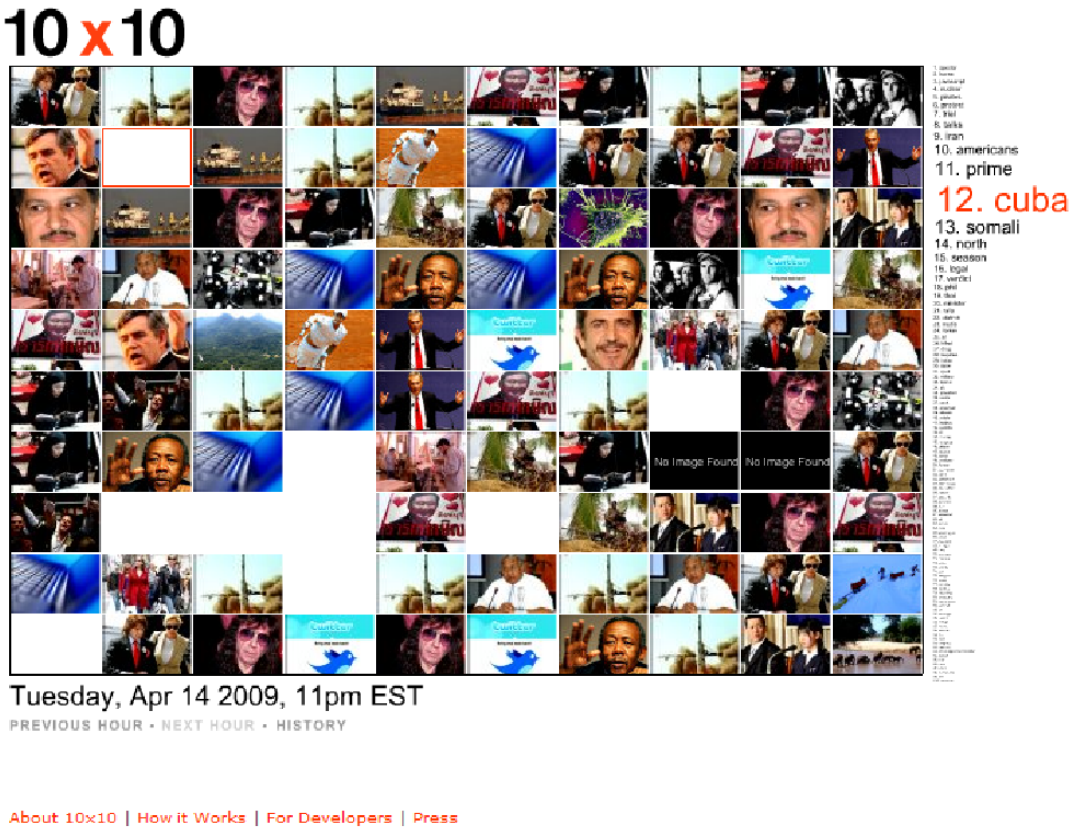

Figura 2.13: Interface do 10 x 10

[10]http://www.tenbyten.org/10x10.html

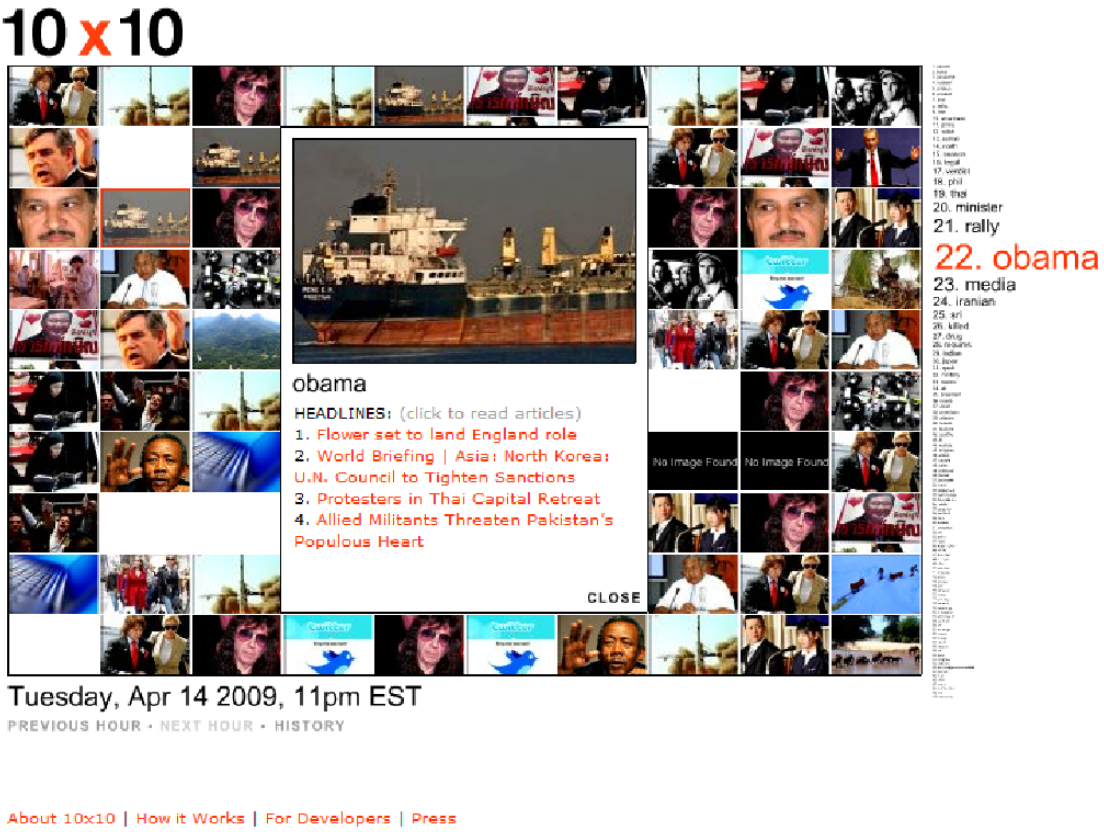

Figura 2.14: Interface do 10 x 10 - Selecção de um acontecimento

2.7.4 We Feel Fine

O We Feel Fine pretende representar o estado de espírito humano à escala mundial. Este projecto é resultado da recolha do estado de espírito de utilizadores de um grande número de blogs (LiveJournal, MSN Spaces, MySpace, Blogger, Flickr, Technorati, Feedster, Ice Rocket, e Google) desde Agosto de 2005. Esta recolha é realizada a partir de frases que sejam constituídas por citações como “I feel“ ou “I am feeling“. Quando uma frase, com estas citações, é encontrada esta é extraída do inicio até ao fim da frase e é guardada na base de dados. De seguida, são identificadas as palavras que expressam sentimento a partir de uma lista de 5000 adjectivos e advérbios, construída manualmente, de forma a identificar o estado de espírito.

Como os blogs são, frequentemente, estruturados segundo um padrão com idade, sexo e localização geográfica do autor, esta informação pode ser extraída e guardada junto com a frase. O resultado é uma base de dados de vários milhões de estados de espírito, que aumenta entre 15.000 e 20.000 novos estados de espírito por dia. Utilizando diferentes interfaces, os estados de espírito podem ser seleccionados e ordenados por diferentes faixas etárias, oferecendo respostas a questões específicas como por exemplo:

1. Os europeus sentem-se tristes com mais frequência do que os americanos?
2. As mulheres sentem-se gordas com mais frequência do que os homens?
3. O tempo chuvoso afecta o nosso estado de espírito?
4. Quais são os sentimentos mais representativos das mulheres nova-iorquinas com 20 anos?
5. O que sentem as pessoas neste momento em Bagdade?
6. Quais foram os sentimentos das pessoas no Dia dos Namorados?
7. Quais são as cidades mais felizes e as mais tristes no mundo?

A interface de resposta a estas questões é representado por um sistema de auto-organização de partículas que representam sentimentos, em que cada partícula de sentimento é obtido de um post. As propriedade das partículas como a cor, tamanho, forma e opacidade representam um sentimento. A partir de um clique na partícula é revelada a frase completa ou fotografia que a partícula representa[11].

2.7.4.1 We Feel Fine - Murmurs

Esta interface apresenta uma lista rolante de estados de espírito por ordem cronológica inversa.

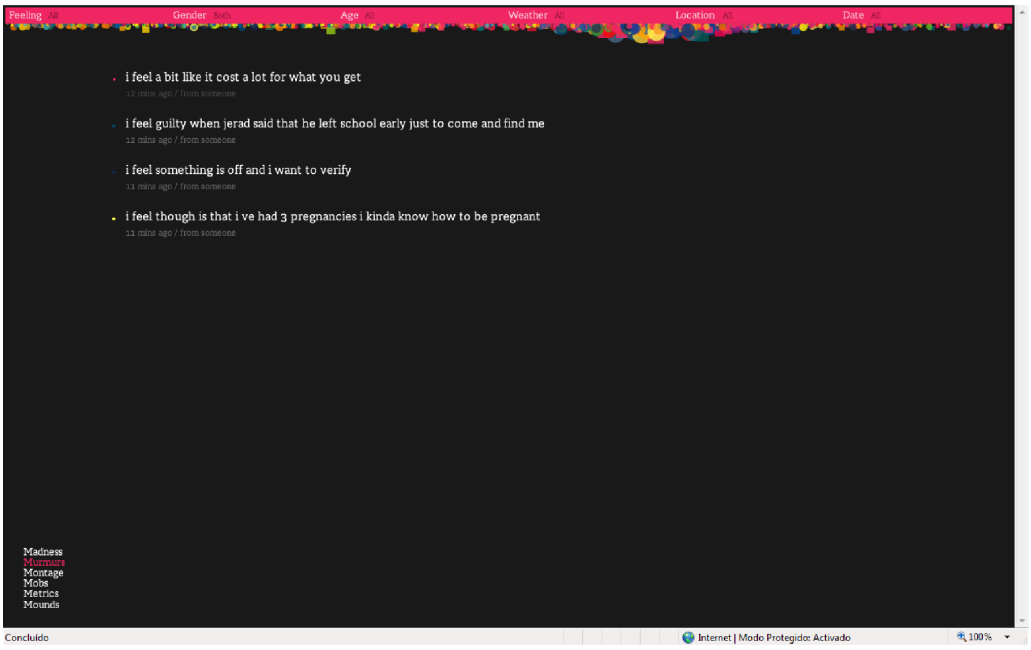

Figura 2.15: Interface Murmurs do We Feel Fine

[11] http://www.wefeelfine.org/

2.7.4.2 We Feel Fine - Montage

A interface Montage apresenta estados de espírito a partir de fotografias disponibilizadas on-line. As fotografias são apresentadas numa grelha 10x10 de tamanho variável, dependendo do número de fotografias disponíveis. Qualquer fotografia da grelha pode ser aumentada, bastando clicar na fotografia.

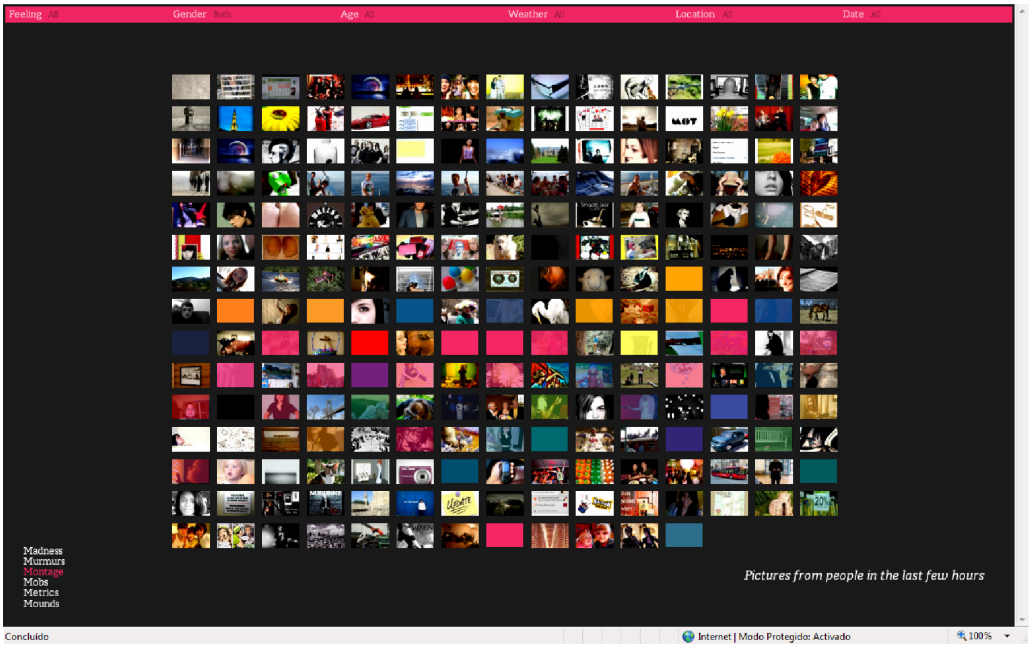

Figura 2.16: Interface Montage do We Feel Fine

2.7.4.3 We Feel Fine - Metrics

A interface Metrics apresenta os sentimentos mais representativos, juntamente com alguns dados estatísticos que explicam o significado dos resultados. Os sentimentos são listados ao longo da margem esquerda da interface. A classificação, representada num círculo vermelho é obtida a partir do número de vezes que a frequência exceda a média global. No lado direito da interface são apresentados gráficos de barras que ilustram o número de vezes que cada estado de espírito ocorreu na amostra da população, bem como o número de vezes que cada estado de espírito normalmente ocorre.

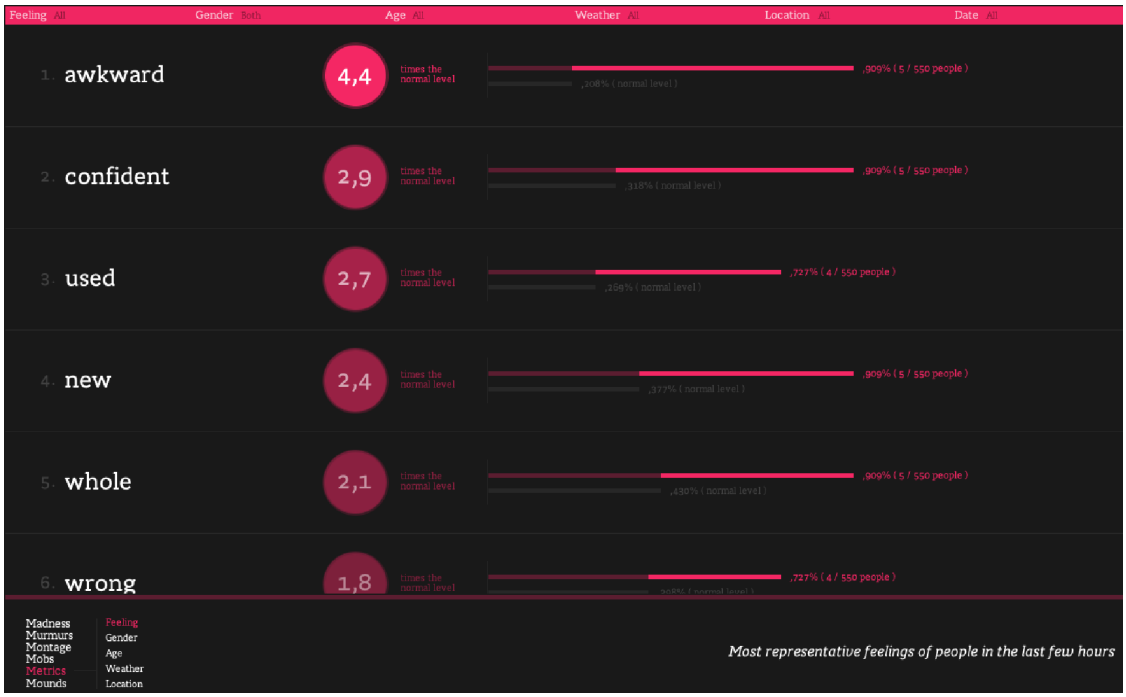

Figura 2.17: Interface Metrics do We Feel Fine

2.7.4.4 We Feel Fine - Mobs

A interface Mobs, apresenta o sentimento mais comum na amostra populacional. Nesta interface, as partículas são ordenadas em linhas por ordem de partilha de estado de espírito. As linhas são ordenadas pelo número de partículas que contêm, e as partículas dentro de cada linha são ordenadas pela duração do período de tempo de cada partícula. As linhas herdam a cor do sentimento que representam. Qualquer partícula pode ser clicada de forma a revelar a frase que expressa o estado de espírito.

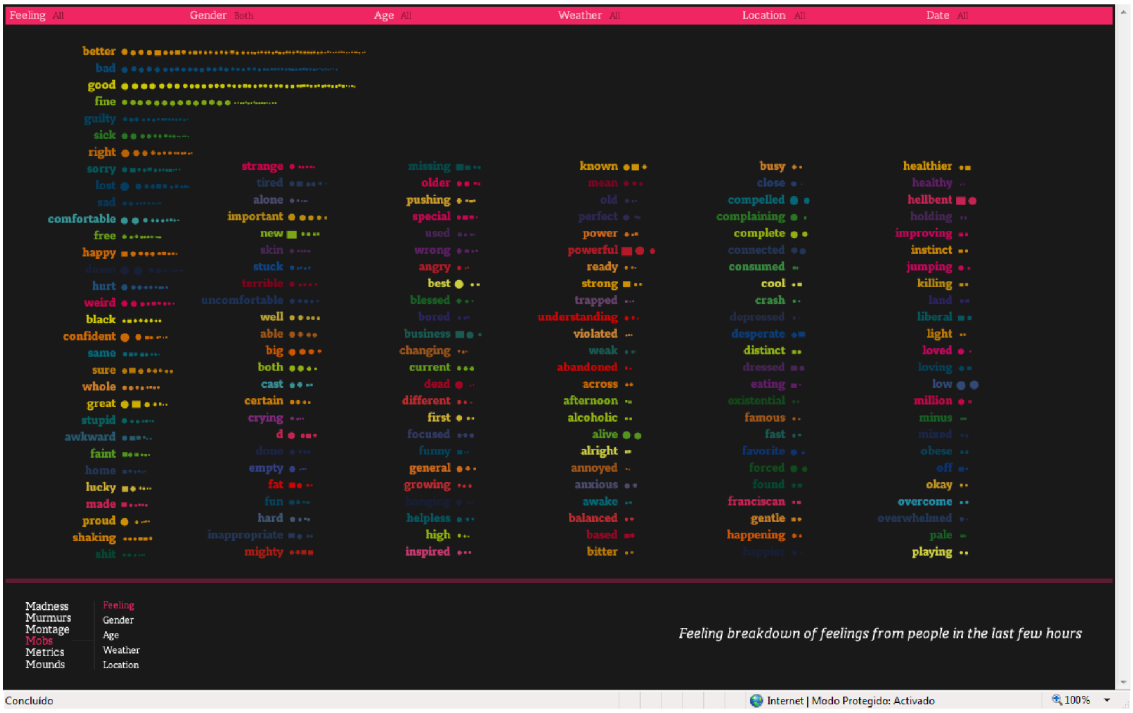

Figura 2.18: Interface Mobs do We Feel Fine

2.7.4.5 We Feel Fine - Mounds

A interface Mounds é independente da população, apresenta cada estado de espírito por ordem de frequência. Cada estado de espírito é retratado por um monte e representado pela cor correspondente ao estado de espírito respectivo.

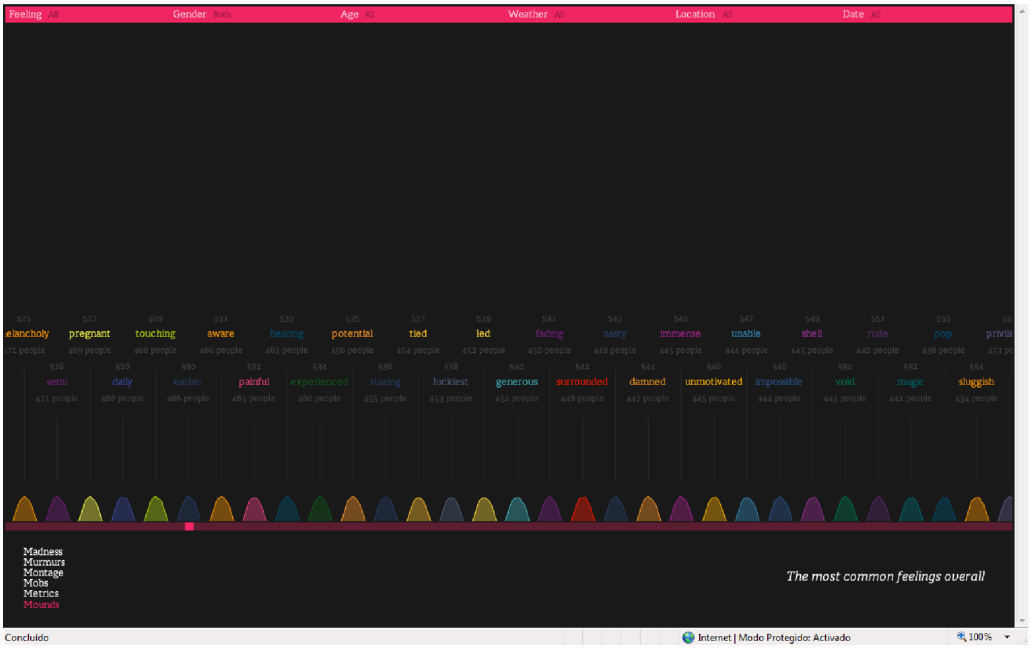

Figura 2.19: Interface Mounds do We Feel Fine

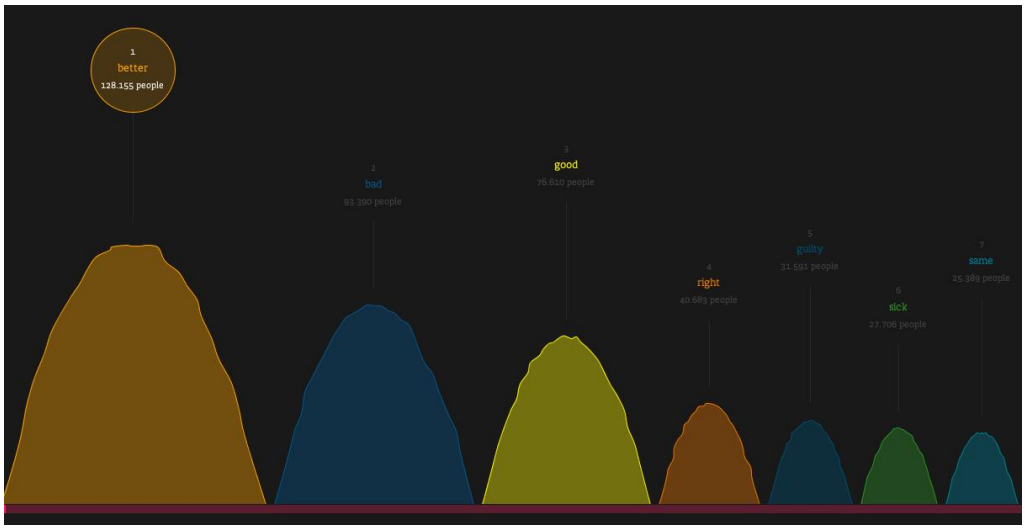

Figura 2.20: Interface Mounds do We Feel Fine - Estados de espírito mais representativos

2.7.5 Social Streams

O Social Streams Live Labs é um projecto da Microsoft Live Labs, cuja missão é pesquisar, agregar e guardar todos os conteúdos de média social. A plataforma Social Streams tem sido utilizada no apoio a diversas aplicações e pesquisas. Uma das aplicações realizada a partir da plataforma Social Streams foi o Political Streams. O Political Streams foi disponibilizado em 2008 nos Estados Unidos no período das Eleições Presidenciais. O objectivo era disponibilizar, de uma forma rápida, as notícias que atraíram maior atenção. A aplicação permite a selecção das entidades, lugares e blogs associados à notícia.[12]

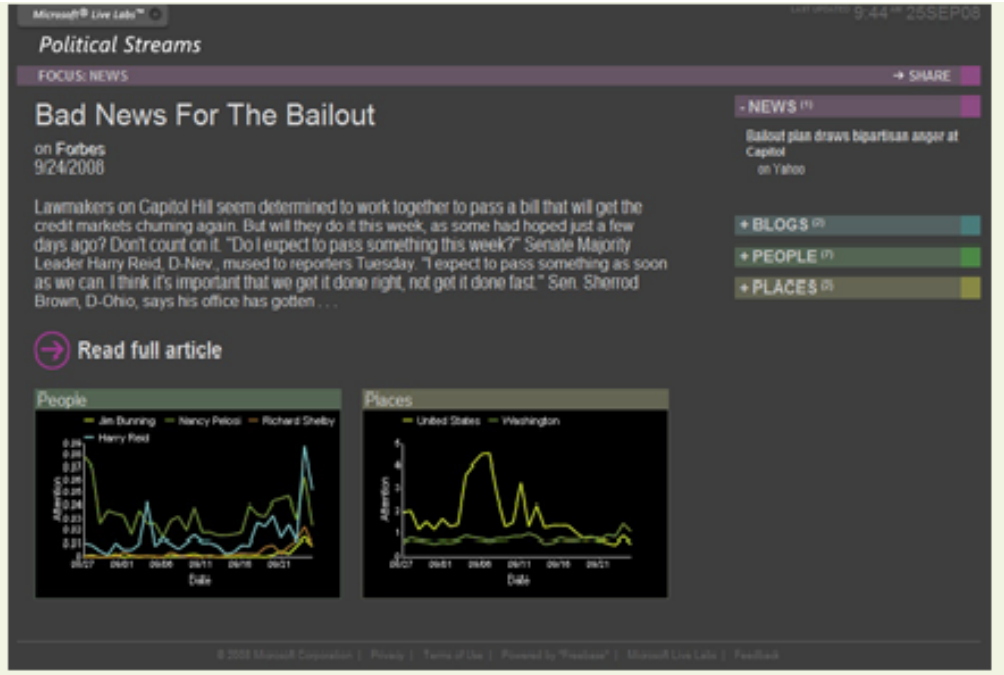

Figura 2.21: Interface do Social Streams

[12] http://livelabs.com/social-streams/

2.7.6 Blews - what the blogosphere tells you about news

O BLEWS da Microsoft Live Labs, utiliza blogs de categoria política para classificar notícias de acordo com a facção, conservadora ou liberal, com que se identifica na blogosfera. Este sistema analisa a informação das notícias que se encontram ligadas a partir de blogs conservadores e liberais e indica o nível de intensidade de discussão de notícias ou assuntos que se encontram na ordem do dia. O BLEWS disponibiliza a funcionalidade *"see the view from the other side"*, que permite ao leitor comparar pontos de vista diferentes sobre a mesma questão a partir de diferentes lados do espectro político. O BLEWS concretiza este objectivo através da análise, em tempo real, dos posts políticos fornecidos pela plataforma Live Labs Social Media, analisando tanto os links como o texto nos blogs. Na interface de visualização actual é apresentada a contagem de links liberais para uma notícia (a azul) e o número de links conservadores (a vermelho). A intensidade da discussão em torno do link / notícia é apresentado como um "indicador de calor"do lado de fora. A intensidade varia de um único quadrado laranja até quatro quadrados brancos . Clicando sobre as "asas"é apresentada uma lista de cada um dos posts com links para as notícias ou artigo. Na lista *dropdown*, as notícias discutidas mais intensamente têm quadrados com contornos esbatidos, enquanto as notícias discutidas normalmente, têm contornos sólidos[13] .

Figura 2.22: Interface do Blews

[13] http://research.microsoft.com/en-us/projects/blews/

2.7.7 EMM NewsExplorer

O NewsExplorer utiliza JRC[14] para gerar automaticamente resumos de notícias, permitindo aos utilizadores consultar as principais notícias de um dia seleccionado, em vários idiomas e comparar a forma como os mesmos eventos são relatados nos meios de comunicação escrita em diferentes línguas. O NewsExplorer permite a selecção automática de noticias a partir de uma lista de entidades mais citadas nas noticias. Esta ferramenta disponibiliza, ainda, outros tipos de selecção, como por exemplo: variantes léxicas dos nomes, títulos, frases, noticias mais recentes e inter-relações entre entidades[15].

Figura 2.23: Interface do News Explorer

[14]O Joint Research Centre (JRC), tem vindo a utilizar tecnologias de análise linguística desde 1998 para combater a sobrecarga de informação e superar as barreira de idioma com a finalidade de apoiar a Comissão Europeia e de instituições dos Estados-Membros. Para esse efeito, foram desenvolvidas ferramentas com capacidade de agregar, analisar e visualizar textos multilingue e fornecer informação cruzada. Estas ferramentas de análise foram integradas com o motor de recolha de notícias (Europe Media Monitor EMM) para produzir várias aplicações, de alto nível.

[15]http://press.jrc.it/NewsExplorer/entities/pt/392388.htm

2.7.7.1 EMM NewsExplorer - Associações entre entidades

Esta ferramenta permite a visualização gráfica das inter-relações "associações" entre pessoas e organizações identificadas nas notícias.

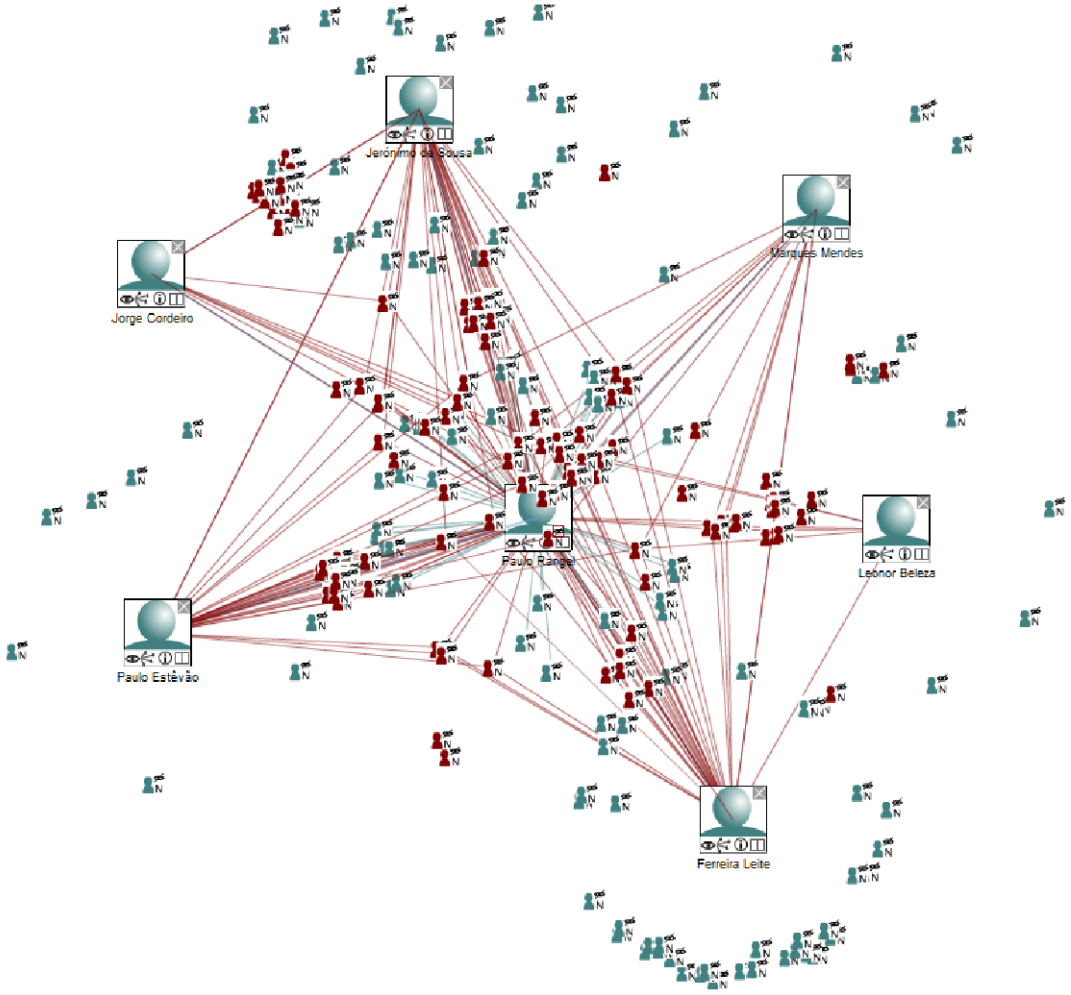

Figura 2.24: Interface do News Explorer - Inter-relacionamento entre entidades

2.7.8 Verbatim

O Verbatim é uma ferramenta de extracção automática de citações dos órgãos de comunicação social portugueses. Esta ferramenta processa diariamente notícias recolhidas da web e funciona sem qualquer intervenção humana. Com esta ferramenta é possível ver o que foi dito por alguém sobre vários tópicos (ex: Teixeira dos Santos ou Barack Obama), tudo o que foi dito sobre um tópico (ex: Crise ou EUA), ou simplesmente saber quais as fonte de uma citação em particular (ex: "disse que o simulacro..."). Esta ferramenta foi desenvolvida por Luís Sarmento (NIAD&R) e Sérgio Nunes (FEUP) no contexto de uma colaboração entre a Universidade do Porto e o SAPO Labs[16].

Figura 2.25: Interface do Verbatim

[16]http://pattie.fe.up.pt/verbatim/

2.7.8.1 Verbatim - Citações de uma Entidade Sobre um Assunto

Nesta interface é possível visualizar citações de personalidades sobre os diversos assuntos e sobre as diferentes personalidades.

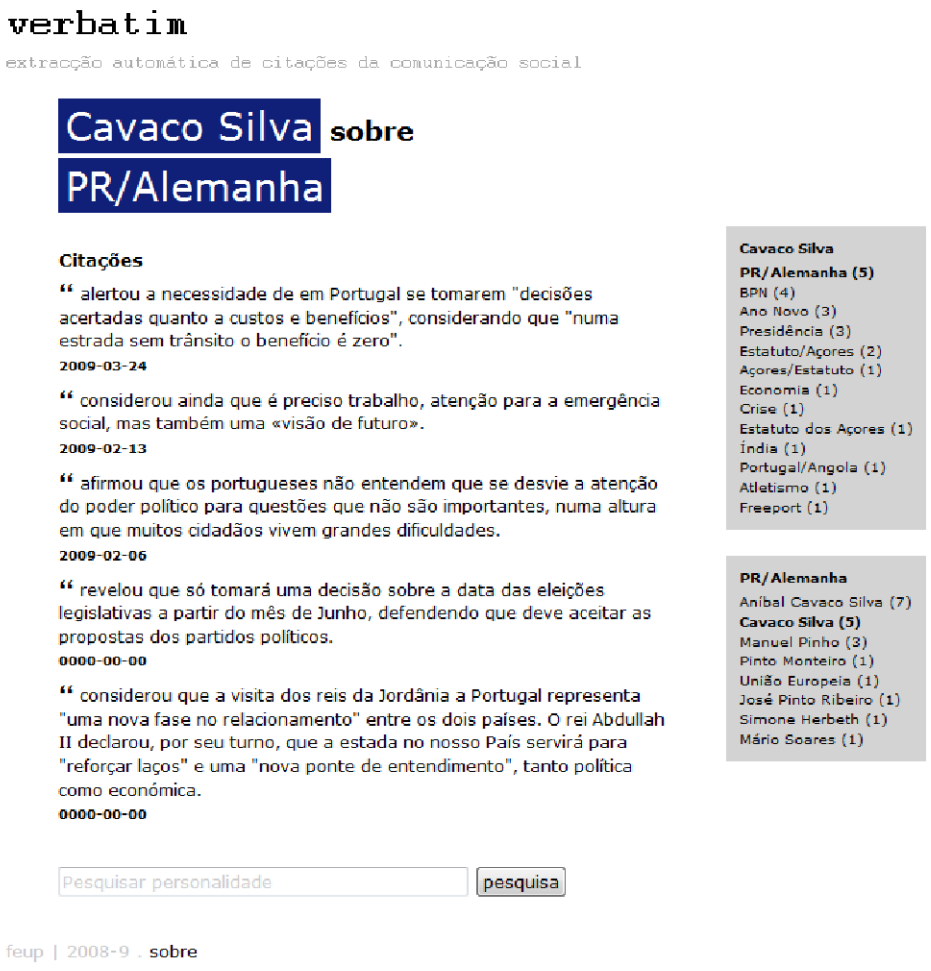

Figura 2.26: Interface do Verbatim - Citações de uma entidade sobre um assunto

2.7.8.2 Verbatim - Visualização de uma Citação de uma Entidade

Exemplo de citações do Presidente da Republica Portuguesa Aníbal Cavaco Silva, sobre diversos assuntos de interesse nacional.

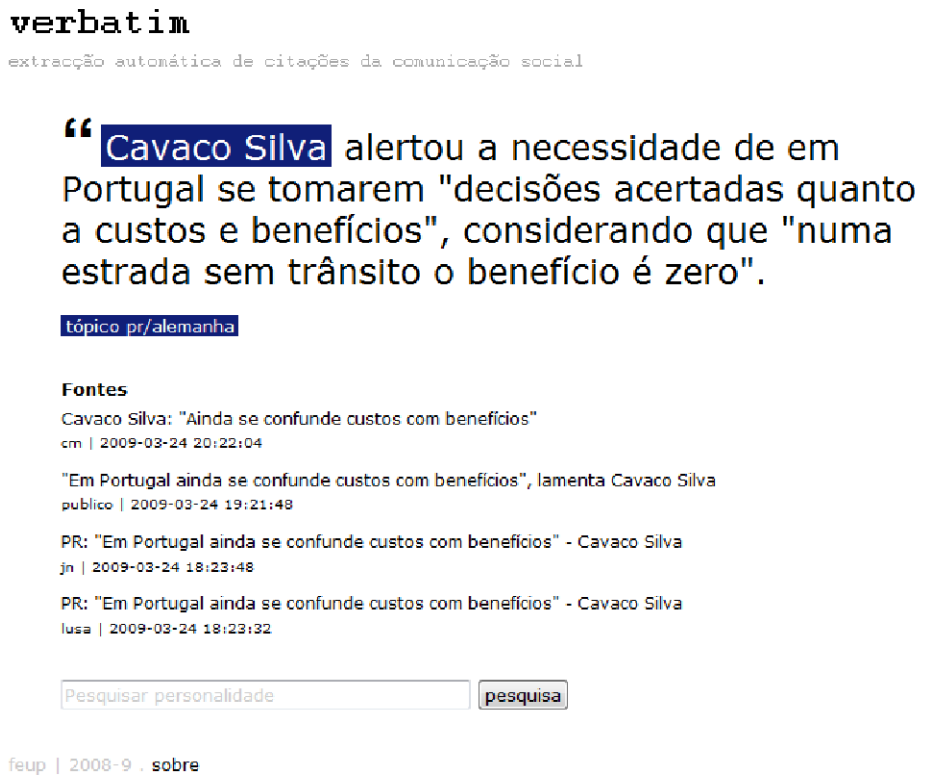

Figura 2.27: Interface do Verbatim - Visualização de uma citação de uma entidade

2.7.8.3 Verbatim ::tendências de personalidades

O Verbatim :: tendências de personalidades, compara as tendências de três personalidades em simultâneo e reflecte estes valores num gráfico de linhas.

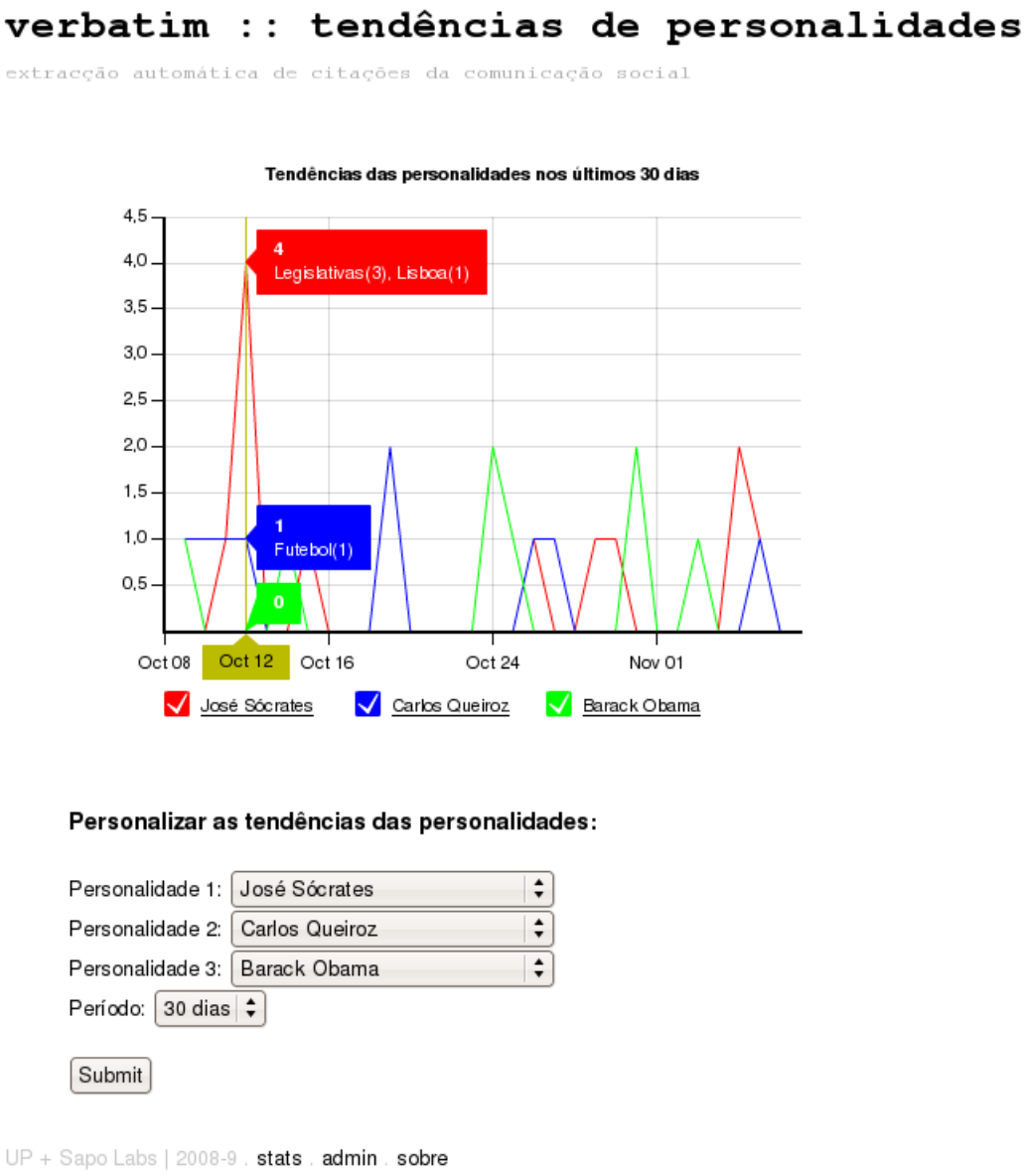

Figura 2.28: Interface do Verbatim :: tendências de personalidades

2.7.8.4 Verbatim :: stats

O Verbatim :: stats permite a visualização, num gráfico de barras, das personalidades e dos tópicos mais activos em períodos de sete dias, trinta dias e um ano.

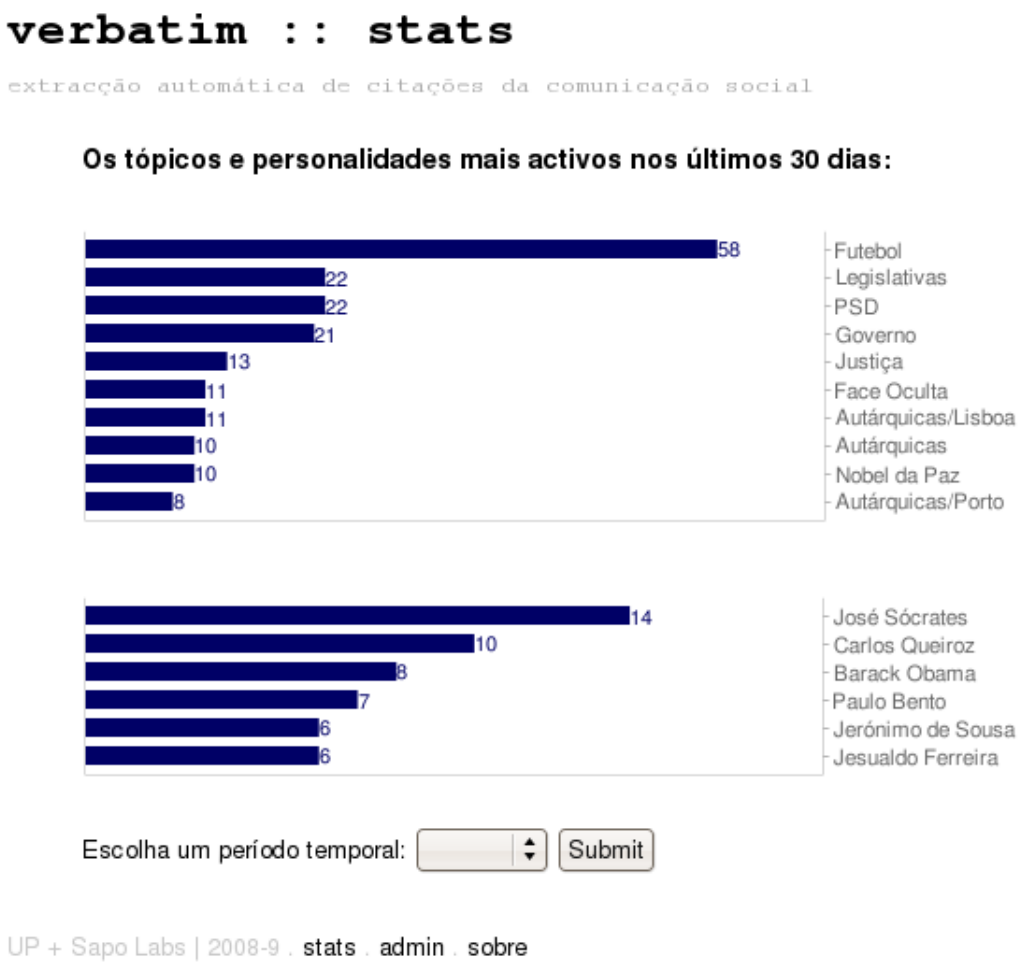

Figura 2.29: Interface do Verbatim :: stats

2.7.9 MemeTracker

O MemeTracker é uma ferramenta que constrói mapas diários de notícias através da análise de cerca de 900.000 notícias e posts, a partir de 1 milhão de fontes on-line.

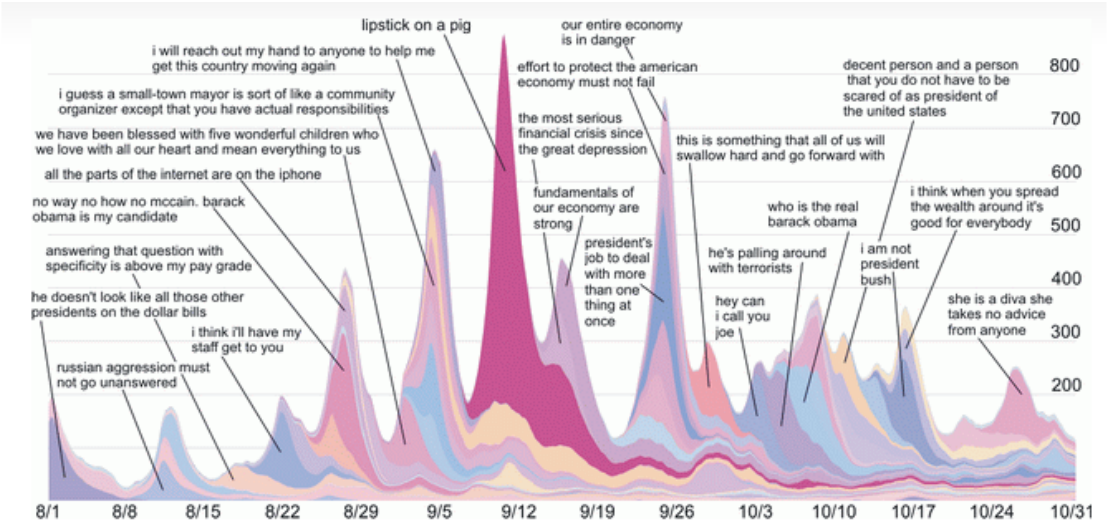

Figura 2.30: Interface do MemeTracker - Com uma notícia ou citação

2.7.9.1 MemeTracker - Selecção de Notícias e Citações

O MemeTracker, selecciona citações e frases que aparecem com maior frequência ao longo do tempo no conjunto de notícias. O MemeTracker possibilita visualizar como as diferentes notícias se apresentam nos blogs e nos sites de notícias e como determinadas notícias persistem enquanto outras desaparecem rapidamente. Globalmente, acompanha mais de 17 milhões de frases e citações diferentes em que cerca de 54% do total das frases / citações mencionadas surgem em blogs e 46% surgem em sites de notícias[17].

Figura 2.31: Interface do Meme Tracker - Citação seleccionada com a respectiva fonte noticiosa

[17] http://memetracker.org/

2.7.10 TextMap

O TextMap referência pessoas, lugares e assuntos que surgem nas notícias, a fim de identificar relações entre estas. O TextMap faz a monitorização do "estado do mundo", analisando tanto a distribuição temporal como espacial destas entidades.

Como entidades são consideradas pessoas, lugares, empresas, cidades, universidades, websites, países, etc. São analisadas, diariamente, mais de 1000 fontes de jornais on-line. Para cada fonte de notícias é mantida a informação numa página que apresenta uma análise gráfica das notícias, por período de tempo e tipo de noticia. O TextMap utiliza técnicas de processamento da linguagem natural para identificar referências a entidades e varias técnicas estatísticas para analisar as justa-posições entre entidades [18].

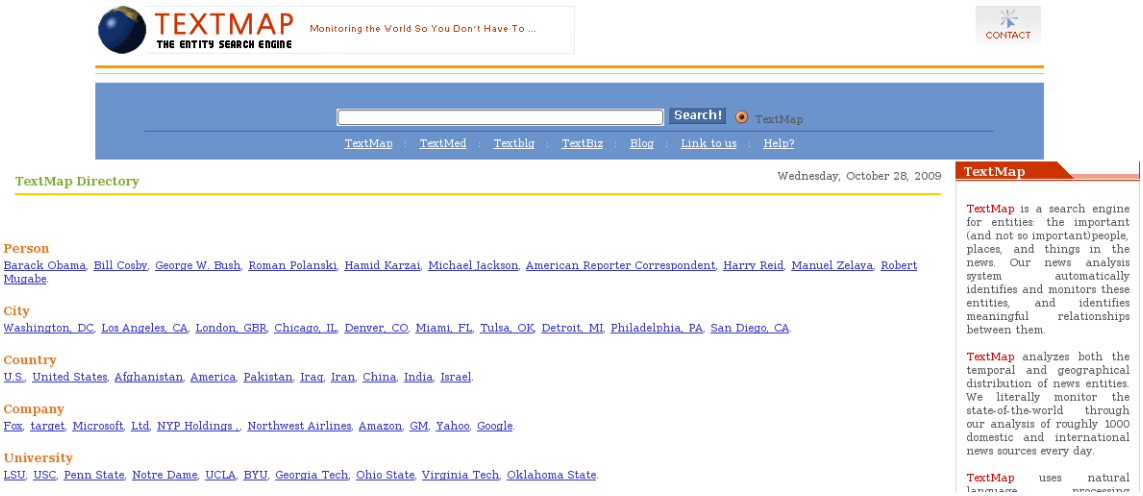

Figura 2.32: Interface do TextMap

[18] http://www.textmap.com/

2.7.10.1 TextMap - Distribuição da Opinião no Mapa Internacional para uma Entidade

Nesta interface é apresentado, num mapa internacional utilizando um código de cores, a distribuição da opinião relativamente à entidade John Edwards no período de 1 a 28 de Março de 2007.

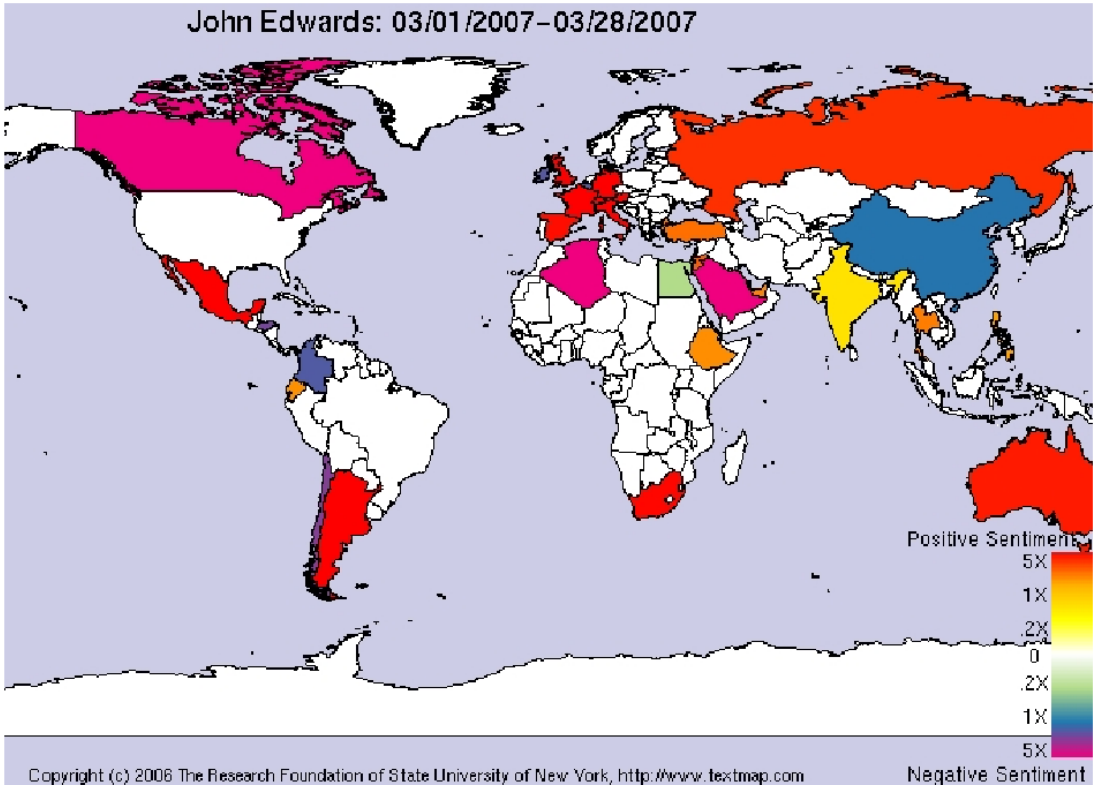

Figura 2.33: Interface do TextMap - Distribuição do sentimento no mapa internacional para uma entidade

2.7.10.2 TextMap - Frequência de Referências à Entidade

Nesta interface é apresentada a frequência de citações, através de um código de cores, representadas num mapa das zonas terrestres com mais citações à entidade John Edwards no período de 1 a 28 de Março de 2007.

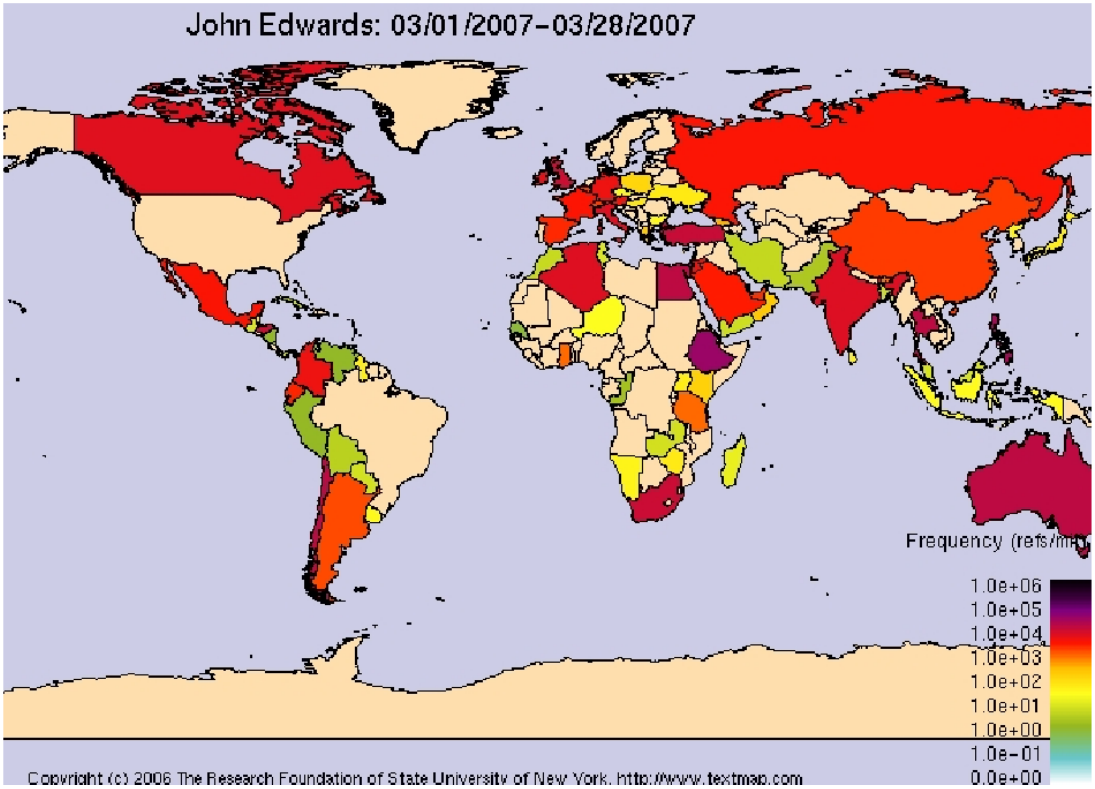

Figura 2.34: Interface do TextMap - Frequência de referências à entidade John Edwards

2.7.10.3 TextMap - Popularidade da Entidade Numa Linha Temporal

Nesta interface é apresentada a popularidade da entidade John Edwards no periodo de 4 de Julho a 29 de Outubro.

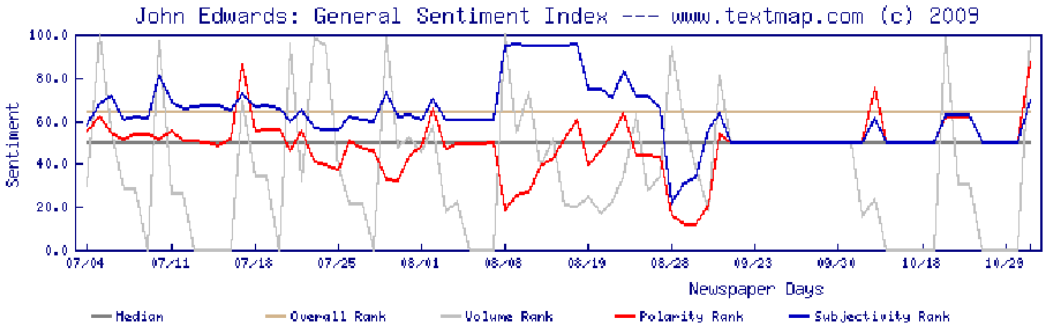

Figura 2.35: Interface TextMap - Popularidade da entidade numa linha temporal

2.7.11 Pollster.com Maps

O Pollster.com analisa e apresenta mapas de resultados de votações efectuadas em diversos sites on-line. Os mapas disponibilizados pelo Pollster.com reflectem as tendências de sondagens realizadas on-line, dos diferentes assuntos que se encontram na ordem do dia. O Pollster.com utiliza as votações realizadas nos sites das diversas fontes noticiosas para reflectir, numa linha temporal, opiniões sobre os assuntos que estão na ordem do dia[19].

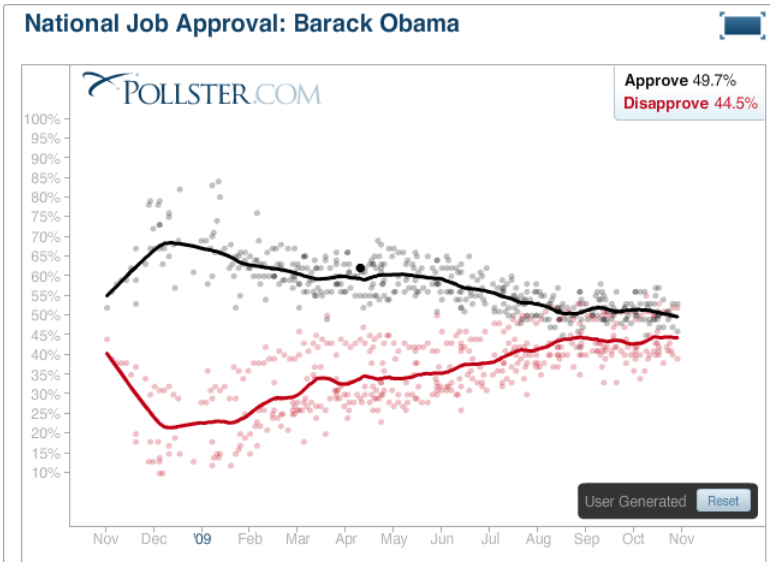

Figura 2.36: Interface do Pollster - Resultados de votações numa linha temporal

Pollster	Dates	N/Pop	Approve	Disapprove	Undecided
Rasmussen	10/30-11/1/09	1500 LV	46	52	-
Gallup	10/30-11/1/09	1500 A	53	39	-
Rasmussen	10/27-29/09	1500 LV	47	52	-
Gallup	10/27-29/09	1500 A	53	40	-
FOX	10/27-28/09	900 RV	50	41	10
YouGov/Polimetrix	10/25-27/09	1000 A	50	43	8
Rasmussen	10/24-26/09	1500 LV	49	51	-
Gallup	10/24-26/09	1500 A	51	41	-
NBC/WSJ	10/22-25/09	1009 A	51	42	7
Rasmussen	10/21-23/09	1500 LV	47	51	-
Gallup	10/21-23/09	1500 A	54	38	-

Figura 2.37: Interface do Pollster - Resultados de votações por fonte noticiosa

[19] http://www.pollster.com/

Figura 2.38: Interface do Pollster - Resultados de votações por Estados

2.8 Resumo

Neste capítulo foram apresentados os dois algoritmos que têm sido utilizados e testados na classificação de texto, nomeadamente Naïve Bayes e Support Vector Machines. Foram estudados diferentes artigos sobre a aplicação destes algoritmos em diferentes tipos de corpus e com objectivos distintos, mas com a finalidade comum de classificação de texto. São também abordados artigos que aplicação diferentes heurísticas na classificação de texto. Em resumo, são apresentados quadros resumo de enquadramento dos diferentes artigos relativamente aos recursos utilizados, técnicas de aprendizagem automática ou heurísticas utilizadas. Nestes quadros resumo são referenciadas as áreas de intervenção desta tese. Por fim, são apresentados exemplos de trabalhos que utilizam estas técnicas e que abordam o tema desta tese.

Capítulo 3

Descrição do Sistema

O sistema desenvolvido é um sistema supervisionado de aprendizagem automática, que analisa e classifica notícias de carácter político, automaticamente, como favoráveis ou desfavoráveis para as entidades que de alguma forma se encontrem referenciadas nas notícias. O sistema faz a recolha, periodicamente, das notícias que são disponibilizadas on-line pelos diversos meios de comunicação social em feeds RSS e agrupa-as numa base de dados especialmente concebida para o efeito. Estas notícias, guardadas na base de dados, constituem o corpus de notícias. Antes de serem guardadas na base de dados, as notícias, passam por um processo automático de extracção de entre as tags HTML e todos os caracteres de marcação de hipertexto, de forma a eliminar qualquer caractere que possa causar ruído no processo de classificação.

O processo seguinte é a identificação e referenciação automática das entidades políticas no corpus criado. Nesta fase é criada uma lista de entidades políticas[1] que irá ser utilizada no processo de identificação automática das entidades em todas as notícia. Posteriormente à identificação de entidades, existe a necessidade de classificar manualmente notícias que serão usadas para treinar e testar o classificador. Esta tarefa de classificação é realizada por humanos e para facilitar, agilizar e diminuir erros na tarefa de anotação foi criada uma ferramenta de anotação manual.

Uma fase especialmente importante é a identificação e geração dos vectores de features. Neste processo são identificados padrões de features que melhor representem as notícias. Para isso foram criados scripts que geram estas features automaticamente, seleccionando e combinam com outros recursos de forma a criar vectores de features consistentes e robustos que permitam treinar o classificador o melhor possível. Depois de geradas as features, treinado o classificador e gerado o modelo SVM, o classificador está pronto a classificar as novas notícias que são recolhidas periodicamente.

[1] Personalidades da políticas nacional e partidos políticos

3.1 Arquitectura do Sistema

Nesta secção é apresentado através da Figura 3.1 a Arquitectura do Sistema em fase de "Treino e Teste" e em fase de "Produção". A fase de treino e teste é constituída pelos seguintes processos:

- Extracção de Notícias - É o processo de obtenção das notícias a partir das suas fontes. Posteriormente são guardadas na base de dados.
- Identificação de Entidades - É o processo de identificação das entidades identificadas nas notícias.
- Anotação - É o processo de anotar manualmente de forma a classificar notícias que servirão para treino do classificador gerando do Modelo SVM posteriormente utilizado na classificação.
- Exemplos de Treino - São os exemplos obtidos da notícia para cada entidade representada nessa notícia, como apresentado na Secção 3.6.3.1.
- Geração de Features - Processo de criação de features a partir das palavras que constituem a notícia, de forma a representar padrões que permitirão a aprendizagem e posterior classificação de notícias.
- Treino do Classificador - Processo através do qual o classificador aprende baseado nos pares exemplo / valor (positivo / negativo) que lhe são submetidos.
- Modelo SVM - Modelo, resultante do processo de treino, que será utilizado para classificação de exemplos de teste.
- Teste - Processo de classificação dos exemplos de teste que, posteriormente, são contabilizados para avaliação da precisão do classificador .

Na fase de produção alguns processos são eliminados, como é o caso do processo de anotação que só serve para a criação de exemplos de treino e teste. É também eliminado o processo de treino visto não haver a necessidade de gerar o modelo SVM.

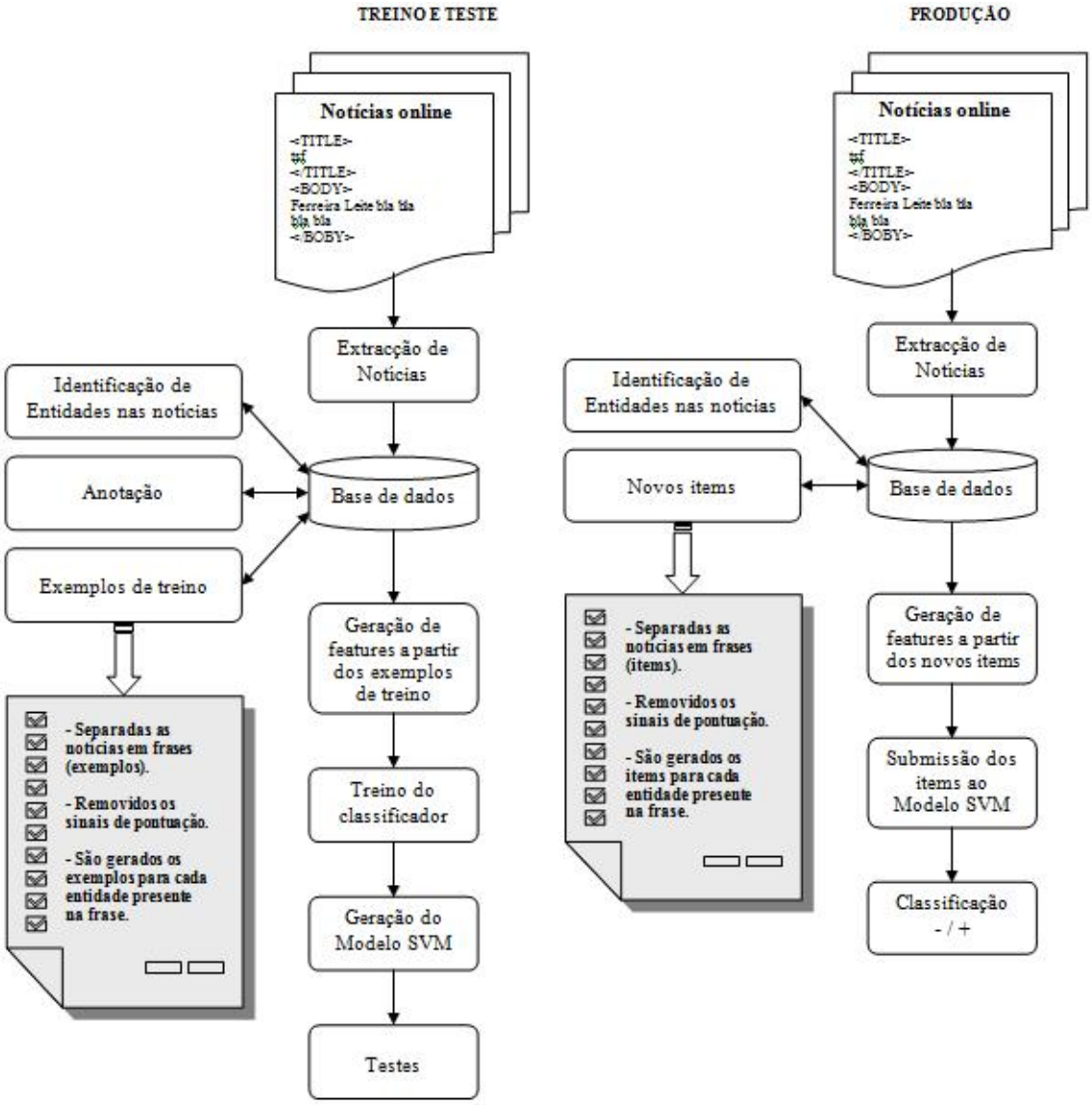

Figura 3.1: Arquitectura dos sistemas de “Treino e Teste” e “Produção”

3.2 Criação do Corpus de Notícias

Na criação do corpus de notícias foram utilizadas diversas fontes nacionais de feeds RSS de notícias, nomeadamente:

- Jornal de Notícias - http://feeds.jn.pt/
- TSF - http://tsf.sapo.pt/
- Expresso - http://aeiou.expresso.pt/
- Público - http://ultimahora.publico.clix.pt/
- Diário de Notícias - http://dn.sapo.pt/

- RTP - http://tv1.rtp.pt/
- SIC - http://sic.aeiou.pt/
- Correio da Manhã - http://www.correiodamanha.pt/
- Visão - http://aeiou.visao.pt/

Foi utilizado uma uma base de dados existente no Laboratório de Inteligência Artificial e Ciência de Computadores da FEUP[2] , utilizada noutros projectos, onde são guardadas as notícias recolhidas periodicamente dos endereços acima referidos. As notícias obtidas a partir dos feeds RSS são constituídas, tipicamente, por um título e um corpo de notícia. O titulo é tipicamente constituído por uma frase e o corpo de notícias é tipicamente constituída por uma, duas ou três frases.

Tabela 3.1: Informação relevante seleccionada de entre as TAGS HTML

Fonte:	Expresso
Categoria:	Política
Data de Publicação:	2008-11-21 20:30:00
Título:	"Sócrates é derrotável já em 2009"
Corpo da Notícia:	Alexandre Relvas faz subir a fasquia de Manuela Ferreira Leite para as próximas legislativas. Em entrevista ao Expresso, o ex-director de campanha de Cavaco Silva e actual presidente do Instituto Sá Carneiro, diz-se "convicto de que Sócrates é derrotável em 2009". E essa será a altura para avaliar a liderança de Manuela Ferreira Leite. Solidário com Manuela, Relvas não exclui em absoluto uma candidatura ao cargo.

Na codificação do script são utilizadas Expressões Regulares para a delimitação e extracção da informação relevante de entre o código fonte obtido do pedido HTTP como no exemplo apresentado na Tabela 3.1.

3.3 Selecção das Entidades

Depois de criado o corpus de notícias, procedeu-se à identificação e criação da lista de entidades. Para esta lista de entidades consideram-se todas as personalidades mais relevantes da política nacional e aquelas personalidades políticas que por uma razão, assunto polémico ou outro assunto que tenha gerado actividade noticiosa, tenham estado na ordem do dia.
As entidades são identificadas pelo seu nome próprio e suas variantes, cargos políticos e

[2]Faculdade de Engenharia da Universidade do Porto

abreviaturas ou expressões que frequentemente referenciam a entidade nas notícias como no exemplo apresentado na Tabela 3.2. Com as entidades identificadas, procedeu-se à selecção das notícias que fazem referência a qualquer uma destas entidades, bem como as entidades presentes e a sua frequência na notícia.

Tabela 3.2: Exemplo de nomes possíveis para entidades

Personalidade: *Manuela Ferreira Leite*	
Nomes possíveis	**Cargos possíveis**
Manuela Ferreira Leite	líder do PSD
Ferreira Leite	presidente do PSD
Personalidade: *José Sócrates*	
Nomes possíveis	**Cargos possíveis**
José Sócrates	Primeiro-Ministro
Sócrates	primeiro-Ministro
	primeiro-ministro

3.4 Ferramenta de Anotação de Notícias

Para agilizar a anotação manual das notícias que são utilizadas para treinar o classificador, foi criada uma ferramenta de marcação de texto. Esta ferramenta, identifica as entidades que são referidas na noticia e demarca-as com cores distintas, facilitando a identificação das mesmas entidades no processo de anotação. As entidades identificadas são listadas e demarcadas com as cores com que se identificam no texto. A cada entidade é atribuído um conjunto de opções, que permite ao anotador seleccionar se a notícia é favorável, desfavorável, neutro ou desconhecido. O período de anotação decorreu entre 20 de Novembro de 2008 e 4 de Maio de 2009. Neste período foram anotadas 3013 notícias utilizadas no treino do classificador.

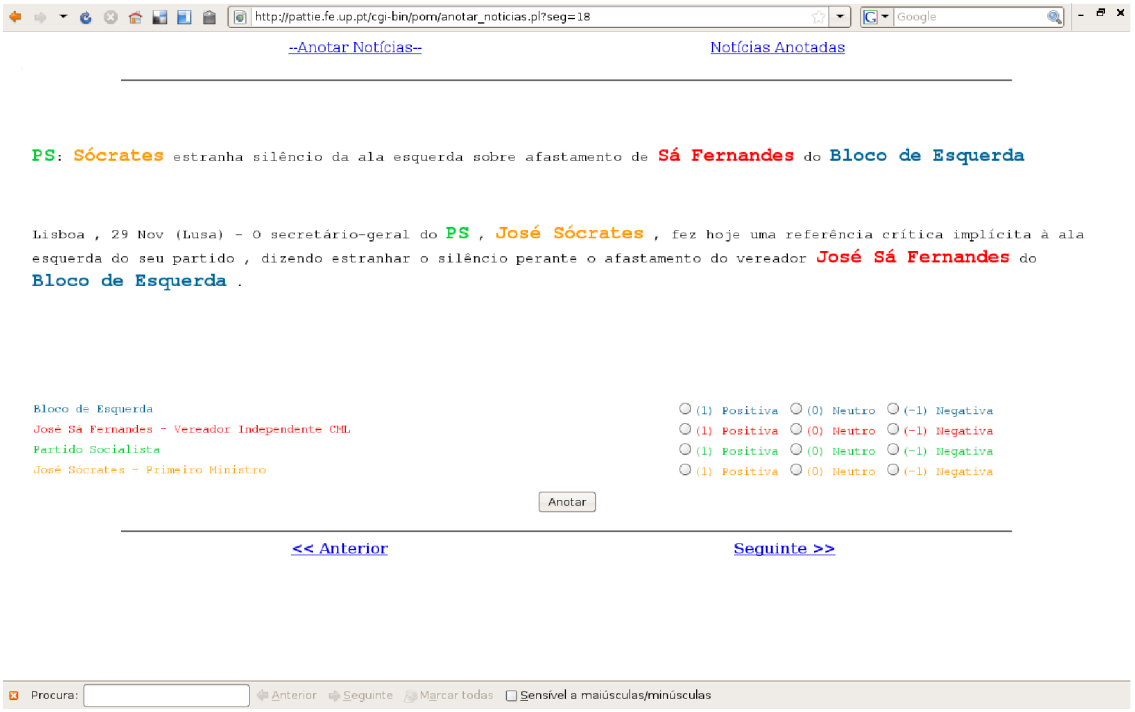

Figura 3.2: Ferramenta de anotação de notícias

3.5 Selecção e Features

Na selecção e criação das features são consideradas apenas as frases da notícia que façam referência às entidades políticas. Destas frases, são geradas as features utilizadas para treinar o classificador SVM.

A notícia é dividida em frases e são extraídos todos os sinais de pontuação. Habitualmente são também extraídas as stopwords, uma vez que estas não representam informação especialmente relevante obtem-se ganho de processamento. No entanto, como optou-se por utilizar SVMs e estas funcionam especialmente bem com exemplos esparsos e com vectores de grandes dimensões e como na classificação de texto existem poucas features irrelevantes [Joa98], optou-se por manter as stopwords. Se considerarmos a notícia apresentada na Tabela 3.1, depois de passar pelo processo acima descrito, esta fica com o aspecto apresentado na Tabela 3.3.

Tabela 3.3: Constituição da notícia

Título:
Sócrates é derrotável já em 2009
Corpo da Notícia:
Alexandre Relvas faz subir a fasquia de Manuela Ferreira Leite para as próximas legislativas
Em entrevista ao Expresso o ex-director de campanha de Cavaco Silva e actual presidente do Instituto Sá Carneiro diz-se convicto de que Sócrates é derrotável em 2009
E essa será a altura para avaliar a liderança de Manuela Ferreira Leite
Solidário com Manuela Relvas não exclui em absoluto uma candidatura ao cargo

3.6 Representação Vectorial

A representação vectorial é o processo de dar configuração à notícia de forma a que SVM consiga identificar padrões nos diversos exemplos de treino que lhe são submetidos. Este processo permite à SVM, comparando com o modelo gerado, classificar novos exemplos. Os exemplos são gerados em função das entidades que foram previamente identificadas na notícia. Cada frase dá origem a um exemplo e cada exemplo dá origem a um vector de features. Na Tabela 3.4 é apresentada uma legenda de etiquetas de features que referenciam as palavras de cada frase em relação à entidade para a qual se está a criar o vector.

Tabela 3.4: Legenda de etiquetas de features

at_palavra (para todas as palavras atrás da entidade)
at3_palavra (3ª palavra atrás da entidade)
at2_palavra (2ª palavra atrás da entidade)
iat_palavra (imediatamente atrás da entidade)
OUTRA_ENTIDADE (outra entidade presente na frase)
iaf_palavra (imediatamente à frente da entidade)
af2_palavra (2ª palavra a seguir à entidade)
af3_palavra (3ª palavra a seguir à entidade)
af_palavra (para todas as palavras à frente da entidade)
conj_af_palavra1_palavra2_... (conjunto de palavras à frente da entidade "pirâmides de palavras")
conj_at_palavra1_palavra2_... (conjunto de palavras atrás da entidade "pirâmides de palavras")
at_LSP_informação_gramatical (informação gramatical das palavras atrás da entidade)
at11_LSPC_informação_gramatical (informação gramatical das palavras atrás da entidade e a posição na frase)
af_LSPC_informação_gramatical (informação gramatical das palavras à frente da entidade)
af1_LSPC_informação_gramatical (informação gramatical das palavras à frente da entidade e a posição na frase)
af_palavra_palavra (conjuntos de bigramas à frente da entidade)
at_palavra_palavra (conjuntos de bigramas atrás da entidade)
atOU_ENT_palavra (palavras atrás de outra entidade)
afOU_ENT_palavra (palavras à frente de outra entidade)
EntreEnt_palavra (palavras que se encontram atrás, à frente e entre outras entidades presente na frase)

3.6.1 Tipos de Features Utilizados

Para representar o vector de features, para cada exemplo, são criados sete tipos de features diferentes:

1. Conjuntos sequenciais de bigramas identificados à frente e atrás da entidade. Dunja Mladenic e Marko Grobelnik [MG98] observam o aumento da precisão em features constituídas por bigramas relativamente a features constituídas por unigramas.

2. Palavras posicionadas na 2ª, 3ª e 4ª posição atrás e a frente da entidade.

3. Conjuntos de features, em pirâmide, sequencialmente formados por palavras posicionadas à frente e atrás da entidade. Johannes Furnkranz, Tom Mitchell e Ellen Riloff [FMR98] observaram que features constituidas por frases aumentam a precisão relativamente a features constituídas por uma única palavra.

4. Palavras identificadas com a posição em relação à entidade.

5. Informação gramatical das palavras e posição em relação à entidade.

6. Palavras identificadas à frente e atrás de outras entidades que se encontrem na frase.

7. Conjuntos de bigramas sequenciais identificadas entre entidades.

3.6.2 Léxico Semântico do Português (LSP)

O LSP é uma base de dados de informação gramatical de palavras em língua portuguesa, desenvolvido no Laboratório de Inteligência Artificial e Ciência de Computadores da FEUP que tem servido de base outros projectos satélite. No projecto que nesta tese é descrito, o LSP fornece a informação gramatical (Lemma, Category, Gender, Number, Radical category e Valency) quando exista, das palavras que constituem as frases das notícias. As palavras da notícia são substituídas pela informação gramatical e as features são criadas após esta substituição. Na Secção 3.6.3.1 é possível verificar um exemplo da aplicação do LSP na criação de features.

Exemplo de resultado obtido directamente do LSP para a palavra “mau”:
Lemma: mau
Category: adj
Gender: m
Number: s
Radical category: adj
Valency: -1

3.6.3 Exemplo de Features Geradas de uma Notícia

Se considerarmos, como exemplo, o título e a primeira frase do corpo da notícia representado na Tabela 3.3, podemos nestas duas frases identificar três entidades, Sócrates, Manuela Ferreira Leite e Alexandre Relvas. Então, é para estas três entidades que são criados os vectores de features e será para estas entidades que pretendemos prever se a notícia é favorável ou desfavorável.

3.6.3.1 Vector de Features em Função das Entidade

Seguidamente é apresentado exemplos de todos os tipos de features utilizados, em função das entidades presentes na frase, como definido na Secção 3.6.1.

Considerando a frase:
Sócrates é derrotável já em 2009

Constituição do vector de features para a entidade ***Sócrates***:

Features do Tipo 1:
af_é_derrotável = 1 af_derrotável_já = 1 af_já_em = 1
af_em_2009 = 1

Features do Tipo 2:
afSec_derrotável = 1 afThird_já = 1 afFourt_em = 1

Features do Tipo 3:
conj_af_é_derrotável = 1 conj_af_é_derrotável_já = 1
conj_af_é_derrotável_já_em = 1
conj_af_é_derrotável_já_em_2009 = 1

Features do Tipo 4:
af_é = 1 af_derrotável = 1 af_já = 1 af_em = 1 af_2009 = 1
iaf_é = 1 af2_derrotável = 1 af3_já = 1 af4_em = 1 af5_2009 = 1

Features do Tipo 5:
af_LSPC_v = 1 af_LSPC_v = 1 af_LSPC_adv = 1 af_LSPC_prep = 1
af1_LSPC_v = 1 af2_LSPC_v = 1 af3_LSPC_adv = 1 af4_LSPC_prep = 1

Os Tipos 6 e 7 não se aplicam a este caso.
Considerando a frase:
Alexandre Relvas faz subir a fasquia de ***Manuela Ferreira Leite*** para as próximas legislativas.

Constituição do vector de features para a entidade ***Alexandre Relvas***:

Features do Tipo 1:
af_faz_subir = 1 af_subir_a = 1 af_a_fasquia = 1 af_fasquia_de = 1
af_de_OUTRA_ENTIDADE = 1 af_OUTRA_ENTIDADE_para = 1 af_para_as = 1
af_as_proximas = 1 af_proximas_legislativas = 1

Features do Tipo 2:
afSec_subir = 1 afThird_a = 1 afFourt_fasquia= 1

Features do Tipo 3:
conj_af_faz_subir = 1
conj_af_faz_subir_a = 1

conj_af_faz_subir_a_fasquia = 1
conj_af_faz_subir_a_fasquia_de = 1
conj_af_faz_subir_a_fasquia_de_OUTRA_ENTIDADE = 1
conj_af_faz_subir_a_fasquia_de_OUTRA_ENTIDADE_para = 1
conj_af_faz_subir_a_fasquia_de_OUTRA_ENTIDADE_para_as = 1
conj_af_faz_subir_a_fasquia_de_OUTRA_ENTIDADE_para_as_proximas = 1
conj_af_faz_subir_a_fasquia_de_OUTRA_ENTIDADE_para_as_proximas
_legislativas = 1

Features do Tipo 4:
af_faz = 1 af_subir = 1 af_a = 1 af_fasquia = 1 af_de
af_OUTRA_ENTIDADE = 1 af_para = 1 af_as = 1 af_proximas = 1
af_legislativas = 1

iaf_faz = 1 af2_subir = 1 af3_a = 1 af4_fasquia = 1 af5_de = 1
af6_OUTRA_ENTIDADE = 1 af7_para = 1 af8_as = 1 af9_proximas = 1
af10_legislativas = 1

Features do Tipo 5:
af_LSPC_v = 1 af_LSPC_v = 1 af_LSPC_prep = 1 af_LSPC_prep = 1
af_LSPC_art = 1 af_LSPC_adj = 1
Features do Tipo 6:
atOU_ENT_faz = 1 afOU_ENT_subir = 1 afOU_ENT_a = 1
atOU_ENT_fasquia = 1 atOU_ENT_de = 1

afOU_ENT_para = 1 afOU_ENT_as = 1 afOU_ENT_proximas = 1
afOU_ENT_legislativas = 1

Features do Tipo 7:
Não se aplica neste caso, uma vez que só se consideram palavras entre duas outras entidades. Aplica-se em frases que tenham pelo menos três entidades.

Constituição do vector de features para a entidade ***Manuela Ferreira Leite***:

Features do Tipo 1:
at_OUTRA_ENTIDADE_faz = 1 at_faz_subir = 1 at_subir_a = 1 at_a_fasquia = 1
at_fasquia_de = 1 af_para_as = 1 af_as_proximas = 1 af_próximas_legislativas = 1

Features do Tipo 2:

afSec_as = 1 afThird_próximas = 1 afFourt_legislativas = 1

Features do Tipo 3:
conj_at_OUTRA_ENTIDADE_faz = 1
conj_at_OUTRA_ENTIDADE_faz_subir = 1
conj_at_OUTRA_ENTIDADE_faz_subir_a = 1
conj_at_OUTRA_ENTIDADE_faz_subir_a_fasquia = 1
conj_at_OUTRA_ENTIDADE_faz_subir_a_fasquia_de = 1
conj_af_para = 1 conj_af_para_as = 1
conj_af_para_as_proximas = 1
conj_af_para_as_proximas_legislativas = 1

Features do Tipo 4:
at_OUTRA_ENTIDADE = 1 at_faz = 1 at_subir = 1 at_a = 1 at_fasquia = 1 at_de = 1

af_para = 1 af_as = 1 af_próximas = 1 af_legislativas = 1
at6_OUTRA_ENTIDADE = 1 at5_faz = 1 at4_subir = 1 at3_a = 1
at2_fasquia = 1 iat_de = 1

iaf_para = 1 af2_as = 1 af2_proximas = 1 af3_legislativas = 1

Features do Tipo 5:
at_LSPC_v = 1 at_LSPC_v = 1 at_LSPC_prep = 1 at_LSPC_prep = 1
af_LSPC_prep = 1 af_LSPC_art = 1 af_LSPC_adj = 1

Features do Tipo 6:
afOU_ENT_faz = 1 afOU_ENT_subir = 1 afOU_ENT_a = 1
afOU_ENT_fasquia = 1 afOU_ENT_de = 1

afOU_ENT_para = 1 afOU_ENT_as = 1 afOU_ENT_proximas = 1
afOU_ENT_legislativas = 1

Features do Tipo 7:
Não se aplica neste caso, uma vez que só se consideram palavras entre duas outras entidades. Aplica-se em frases que tenham pelo menos três entidades.

Por este processo de geração de features e criação dos respectivos vectores, passam tanto os exemplos [3] obtidos das notícia para treino como os items para predição.

[3] frases da notícia

3.7 Support Vector Machines

Considerando a Figura 3.3 cujas features das palavras (positivas a verde e negativas a vermelho) na figura representam os eixo. Ao submeter à classificação o exemplo "Sócrates é derrotável já em 2009", uma aproximação ao plano encontrado pela SVM seria o plano que se encontra representado na figura. O resultado obtido na classificação deste exemplo seria o valor do comprimento do vector $\vec{d}$ perpendicular ao plano encontrado. Na Figura 2.6 é apresentada uma representação a duas dimensões deste processo. O valor real de classificação deste exemplo é apresentado na Tabela 4.2 do Capitulo 4.

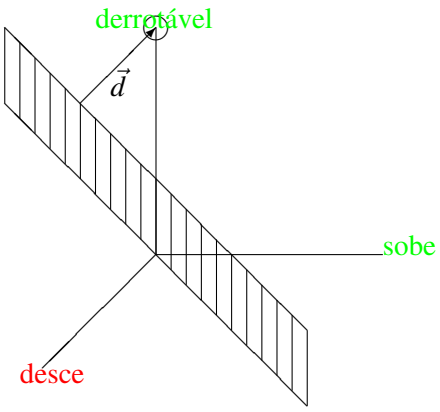

Figura 3.3: Representação das SVMs na classificação de texto

Na implementação do classificador apresentado nesta dissertação foi utilizado o módulo de software SVMLight de Thorsten Joachims [4] que implementa algoritmos para desenvolvimento de sistemas de aprendizagem automática e reconhecimento de padrões. O tipo de kernel utilizado foi kernel linear.

[4]Professor Associado no Departamento de Ciência da Computação da Universidade de Cornell - http://svmlight.joachims.org/

3.8 Dificuldades

Uma situação de difícil resolução por qualquer classificador é realmente as notícias em que existam expressões irónicas. Situações destas podem induzir em erro o classificador.

A notícia que seguidamente se apresenta, é um exemplo de expressão irónica que muitas vezes surgem publicados e que fazem parte do discurso político.

"PSD: "Sócrates no país das maravilhas"
Líder parlamentar do PSD , Paulo Rangel , acusou o primeiro-ministro , José Sócrates , de ter feito no debate do Estado da Nação anúncios de milhões de euros que não pode cumprir."

3.9 Resumo

Neste capítulo foi apresentado e descrito os processos da arquitectura do sistema de "Treino e Teste" e da arquitectura do sistema em "Produção". Foram enumeradas as fontes de notícias e explicado o processo de extracção de informação relevante que constitui o corpus de notícias. Descreveu-se a lista de entidades e a sua identificação na notícia. Foi apresentada a ferramenta de anotação de notícias e descrito o seu objectivo e modo de funcionamento. Descreveram-se os tipos de features e apresentaram-se recursos utilizados na criação das features. Para finalizar é realizada uma aproximação à forma de classificação das SVMs. Importa reter o processo de criação de features, uma vez que o comportamento destas será analisado nos próximos capítulos.

Capítulo 4

Avaliação

Neste capítulo são apresentadas as diferentes métricas utilizadas na avaliação dos diferentes tipos de features, seguindo normas standard [Lew95]. É descrito o método de selecção de exemplos nas notícias e apresentado o esquema K-fold Cross Validation utilizado. São apresentados os resultados obtidos da submissão à classificação de um exemplo demonstrativo.

4.1 Exemplos e Métodos Utilizados na Avaliação

Os exemplos utilizados, para treino e teste nos diversos tipos de avaliação, são obtidos das notícias on-line e separados por frases. As frases da notícia que se consideram, são apenas aquelas em que se identifica uma entidade política (José Sócrates, Manuela Ferreira Leite, PSD, ...) e com uma dimensão superior a 20 caracteres. A base de exemplos é constituída no total por 3918 exemplos equilibrados, nomeadamente 1959 positivos e 1959 negativos. Esta base de exemplos é obtida a partir da anotação manual, realizada a partir da ferramenta de anotação de notícias propositadamente desenvolvida para esta finalidade. A interface de anotação desta ferramenta é apresentada na Secção 3.4.

4.1.1 Precisão

A Precisão é calculada a partir da fracção entre o número de exemplos que o classificador acertou em relação ao número de exemplos testados [MKSW99].

$$\text{Precisão} = \frac{|\{\text{Exemplos certos}\} \bigcap \{\text{Exemplos testados}\}|}{|\{\text{Exemplos testados}\}|}$$

4.1.2 Abrangência

A Abrangência é calculada a partir da fracção entre o número de exemplos que o classificador acertou e o número total de exemplos [MKSW99].

$$\text{Abrangência} = \frac{|\{\text{Exemplos certos}\} \bigcap \{\text{Total dos exemplos}\}|}{|\{\text{Total dos exemplos}\}|}$$

4.1.3 K-fold Cross Validation

Na avaliação K-fold Cross Validation uma amostra original é dividida em K sub-amostras. Destas K sub-amostras é retirado uma para teste e as restantes K-1 sub-amostras são utilizadas para treinar o modelo. Este processo é então repetido K vezes e os resultados combinados de forma a obter-se uma única estimativa. O valor 10 para K é o valor utilizado com mais frequência [Koh95].

No nosso esquema de avaliação K-fold Cross Validation, do conjunto total de 3918 exemplos foram seleccionados aleatoriamente 90% para treino, que resultou em aproximadamente 3526 exemplos e testaram-se os restantes 10%, aproximadamente 392 exemplos. Este processo foi realizado 10 vezes, garantindo-se que os exemplos de teste nunca se repetem. Assim, todos os exemplos foram utilizados para treinar o classificador e submetidos à classificação.

4.1.4 Avaliação das Features Precisão vs Abrangência

Na avaliação das features Precisão vs Abrangência é utilizado o esquema K-fold Cross Validation anteriormente apresentado. No final é calculada a média aritmética dos resultados certos e dos resultados errados. Com estes valores médios é calculada a Precisão e a Abrangência.

4.2 Exemplo de Classificação

Vamos considerar a notícia representada na Tabela 3.1 como exemplo. No processo de classificação a notícia é dividida em frases em que se identifiquem entidades. Cada

entidade, identificada nas diferentes frases, dá origem a um exemplo, como apresentado na Tabelas 4.1. Cada exemplo destes passa pelo processo de geração de features, dando origem a um vector de features que é submetido à classificação. O resultado desta classificação é apresentado na Tabelas 4.2. Dos resultados obtidos podemos verificar que a notícia para Alexandre Relvas e para Cavaco Silva é positiva, para José Sócrates é negativa. No caso dos resultados obtidos para a Manuela Ferreira Leite, a polaridade da notícia para esta entidade vai depender do método escolhido para desambiguar situações em que existam exemplos positivos e exemplos negativos para a mesma entidade. Neste projecto optou-se pela maior frequência de polaridade dos exemplos da notícia (positivos / negativos). Segundo este método a notícia é negativa para Manuela Ferreira Leite, uma vez que dois exemplos foram classificados como negativos e um como positivo.

Tabela 4.1: Exemplos para teste obtidos de uma notícia

Entidade	**Exemplo**
Alexandre Relvas	faz subir a fasquia de OUTRA_ENTIDADE para as próximas legislativas Solidário com OUTRA_ENTIDADE não exclui em absoluto uma candidatura ao cargo
Ferreira Leite	OUTRA_ENTIDADE faz subir a fasquia de para as próximas legislativas E essa será a altura para avaliar a liderança de Solidário com OUTRA_ENTIDADE não exclui em absoluto uma candidatura ao cargo
Cavaco Silva	Em entrevista ao Expresso o ex-director de campanha de e actual presidente do Instituto Sá Carneiro diz-se convicto de que OUTRA_ENTIDADE é derrotável em 2009
José Sócrates	é derrotável já em 2009 Em entrevista ao Expresso o ex-director de campanha de OUTRA_ENTIDADE e actual presidente do Instituto Sá Carneiro diz-se convicto de que é derrotável já em 2009

Tabela 4.2: Resultado da classificação dos exemplos

Entidade	**Exemplo**
Alexandre Relvas	0.99 0.37
Ferreira Leite	-0.03 -0.39 0.76
Cavaco Silva	0.68
José Sócrates	-0.11 -0.90

4.3 Resumo

Neste capítulo foram apresentadas as métricas standard normalmente aplicadas na avaliação deste tipo de classificadores. Foram apresentados exemplos de teste, cujas features foram submetidas à classificação e foram apresentados os resultados da classificação. Importa manter presente os exemplos apresentados, uma vez que são as features de exemplos como este que, no capítulo seguinte, merecerão especial atenção.

Capítulo 5

Análise de Desempenho

Neste capítulo é apresentada a análise ao desempenho do classificador do ponto de vista da Precisão em função da Abrangência e do ponto de vista da finalidade a que se propunha.

É apresentada a análise ao desempenho de cada tipo de features utilizado no classificador bem como ao conjunto das features.

Por considerarmos as Eleições Europeias um bom ponto de controlo e por os resultados das sondagem ficarem longe da realidade, que atribuía ao PS a vitória, efectua-se a análise de valores acumulados da classificação de notícias em relação aos candidatos às Eleições Europeias do dia 7 Junho de 2009 e respectivos partidos políticos. Esta análise pretende avaliar a possibilidade de utilização dos valores acumulados ao longo do tempo na detecção de tendências eleitorais. A avaliação será realizada comparando as tendências apresentadas pelos valores acumulados, relativamente à polaridade das notícias que seria de esperar no dia seguinte às eleições. Esta avaliação foi reforçada com a avaliação manual das notícias desse dia que confirma os resultados.

5.1 Análise aos Resultados da Precisão vs Abrangência dos Diversos Tipos de Features Utilizados

Nesta secção são analisadas as curvas Precisão em função da Abrangência, para todos os tipos de features e para o conjunto das features. São realizadas observações aos pontos fortes e fracos por tipo de features e são avaliados eventuais riscos na utilização de alguns tipos de features.

Ao observar as curvas do gráfico da Figura 5.1 podemos constatar que em todas elas, à medida que a Abrangência aumenta a Precisão diminui. No entanto, todas as curvas têm declínios distintos. Em algumas curvas podemos verificar que existe diminuição rápida da

Precisão em relação à Abrangência, noutros casos a Precisão mantem-se elevada, muito próximo dos 80%. Em algumas curvas é possível verificar que as features não obtêm classificação para valores de Abrangência entre 0 e aproximadamente 0.2, o que torna a utilização destas features arriscado para estes valores de Abrangência.

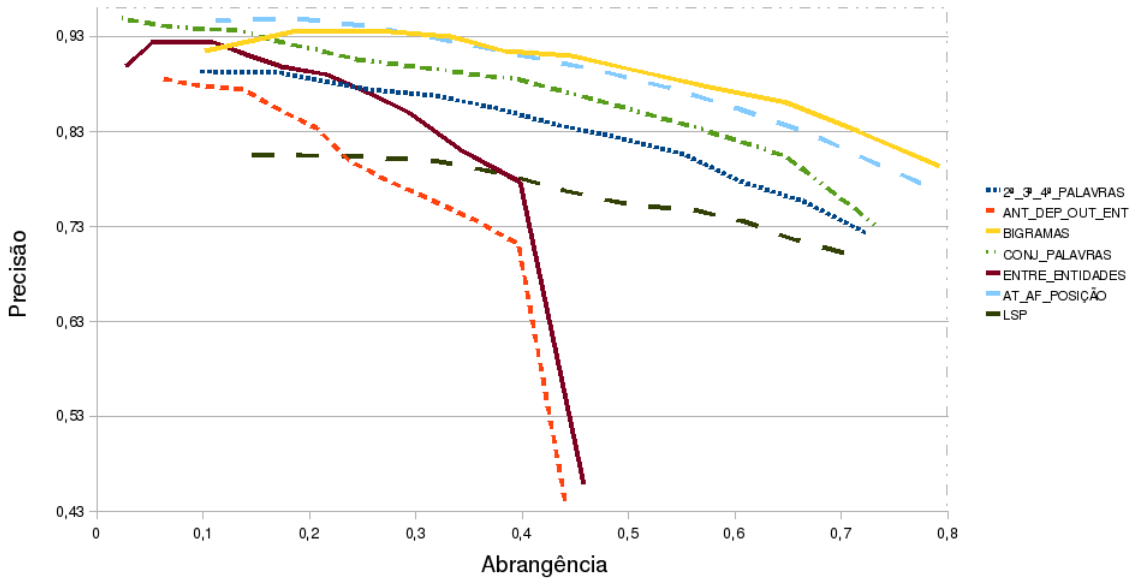

Figura 5.1: Gráfico Precisão vs Abrangência de todas as Features

Precisão vs Abrangência das Features Constituídas por Palavras Posicionadas à Frente e Atrás e Posição em Relação à Entidade em Causa

Ao observar a curva do gráfico representado na Figura 5.2 verifica-se que para uma Abrangência entre 0.1 e aproximadamente 0.2 a Precisão sobe ligeiramente até 0.95. A partir de 0.2 de Abrangência até 0.7 observa-se um decréscimo uniforme na Precisão fixando-se nos 0.78. Estas features demonstram serem boas, uma vez que mantêm a Precisão elevada por toda a Abrangência. A elevada Precisão conseguida por estas features permite observar que a posição das palavras em relação à entidade é muito importante para a composição das features.

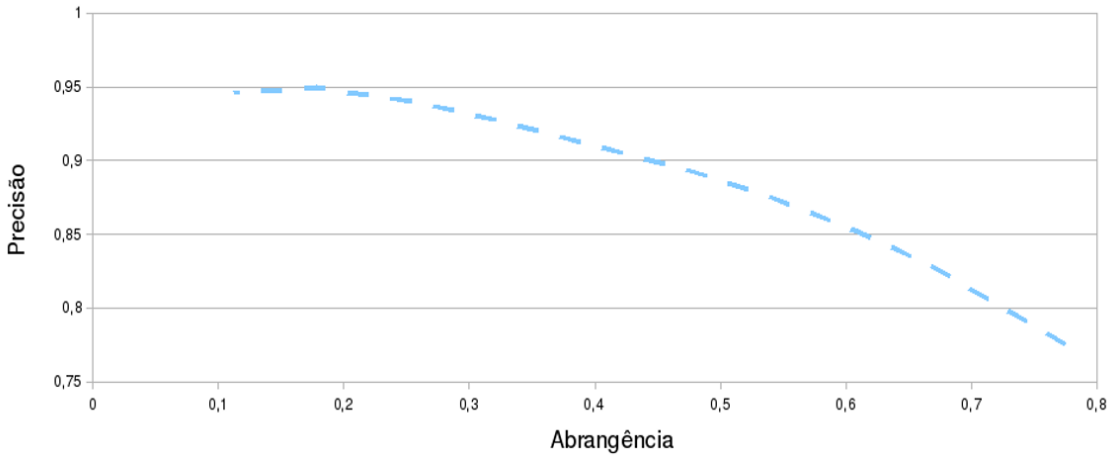

Figura 5.2: Precisão vs Abrangência das Features constituídas por palavras posicionadas à frente e atrás e posição em relação à entidade em causa

Precisão vs Abrangência das Features Constituídas por Palavras Posicionadas à Frente e Atrás de Outras Entidades

A curva representada no gráfico da Figura 5.3 apresenta uma relação Precisão vs Abrangência que manifesta um comportamento decrescente, acentuando-se significativamente a partir dos 0.4 de Abrangência. Para a baixa Precisão destas features contribui o facto de nem em todas as notícias se identificarem mais de uma entidade. Razão esta que, por si só, reduz a Precisão uma vez que uma boa parte das notícias fazem referência a uma única entidade.

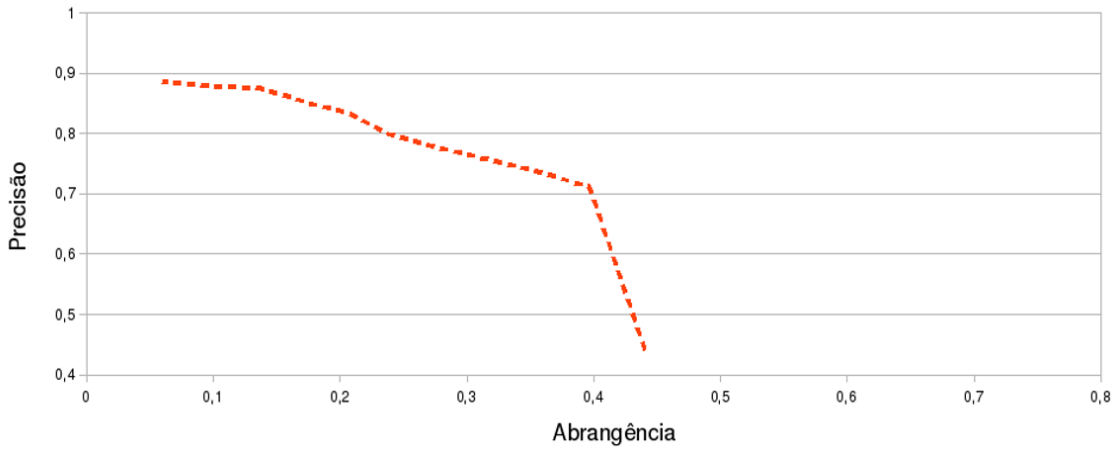

Figura 5.3: Precisão vs Abrangência das Features constituídas por palavras posicionadas à frente e atrás de outras entidades

Precisão vs Abrangência das Features Constituídas Pelas 2ª, 3ª e 4ª Palavras à Frente e Atrás da Entidade em Causa

A curva representada no gráfico da Figura 5.4 apresenta uma Precisão de aproximadamente 0.9 até uma Abrangência de 0.2, descrevendo a partir deste valor um decréscimo uniforme da Precisão, fixado-se em aproximadamente 0.73.

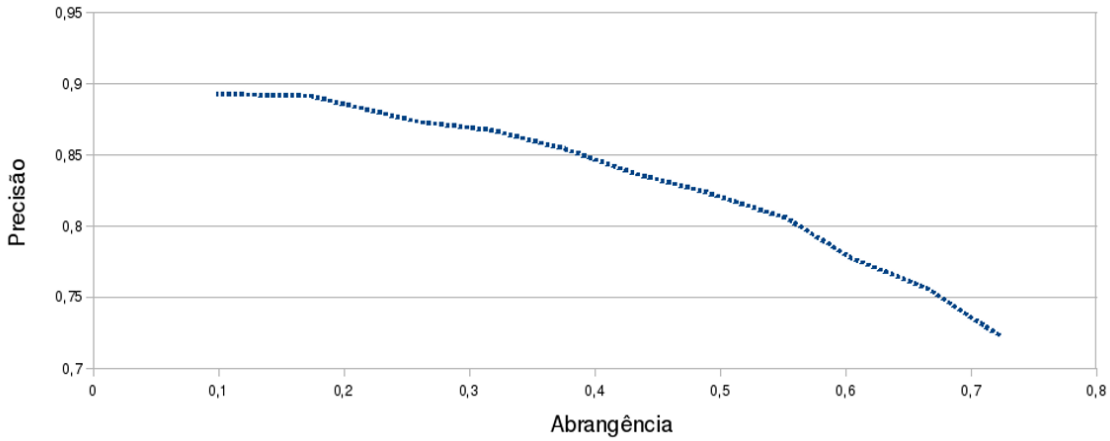

Figura 5.4: Precisão vs Abrangência das Features constituídas pelas 2ª, 3ª e 4ª palavras à frente e atrás da entidade em causa

Precisão vs Abrangência de Features Constituídas por Combinações de Bigramas Posicionados à Frente e Atrás da Entidade em Causa

O comportamento da curva representada no gráfico da Figura 5.5 assemelha-se muito à curva representada no gráfico da Figura 5.2, no entanto, nesta curva para uma Abrangência baixa na ordem dos 0.2 apresenta um decréscimo na Precisão. Na curva da Figura 5.5 para uma Abrangência entre 0.1 e 0.3 apresenta exactamente o comportamento inverso incrementando a Precisão. A Precisão consegue manter-se elevada, acima de 0.93, para uma Abrangência entre 0.2 e 0.4. Estas features demonstram conseguir manter a Precisão bastante elevada por toda a Abrangência. Do comportamento destas features podemos concluir, que features constituídas simplesmente por combinações de palavras conseguem Precisões bastante elevadas.

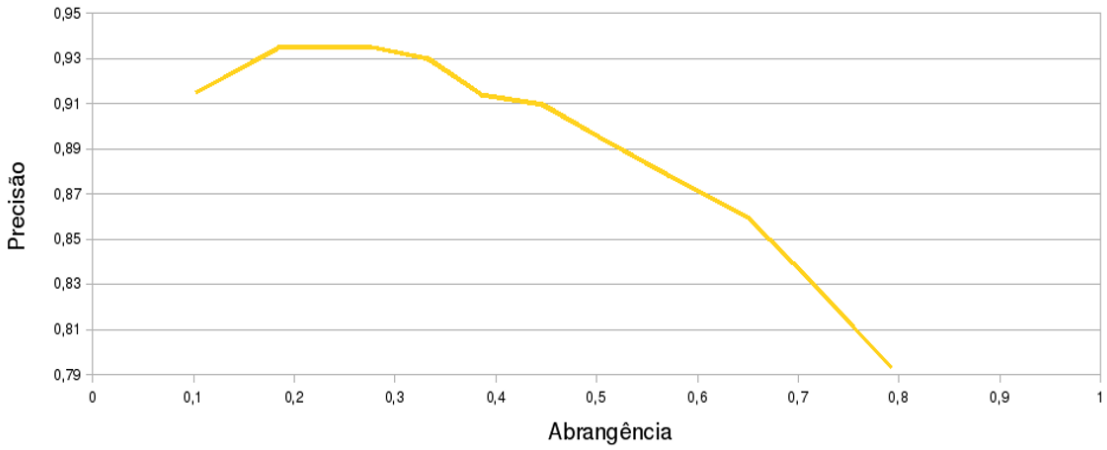

Figura 5.5: Precisão vs Abrangência de Features constituídas por combinações de bigramas posicionados à frente e atrás da entidade em causa

Precisão vs Abrangência de Features Constituídas por Pirâmides de Palavras Posicionadas Consecutivamente à Frente e Atrás da Entidade em Causa

A curva representada no gráfico da Figura 5.6, apresenta um decréscimo praticamente constantes, fixando a Precisão em 0.73. Estas features conseguem obter uma Precisão de aproximadamente 0.93 para uma Abrangência de 0.2. Uma conclusão que se poderá tirar destes resultados é que features constituídas por frases conseguem bom desempenho.

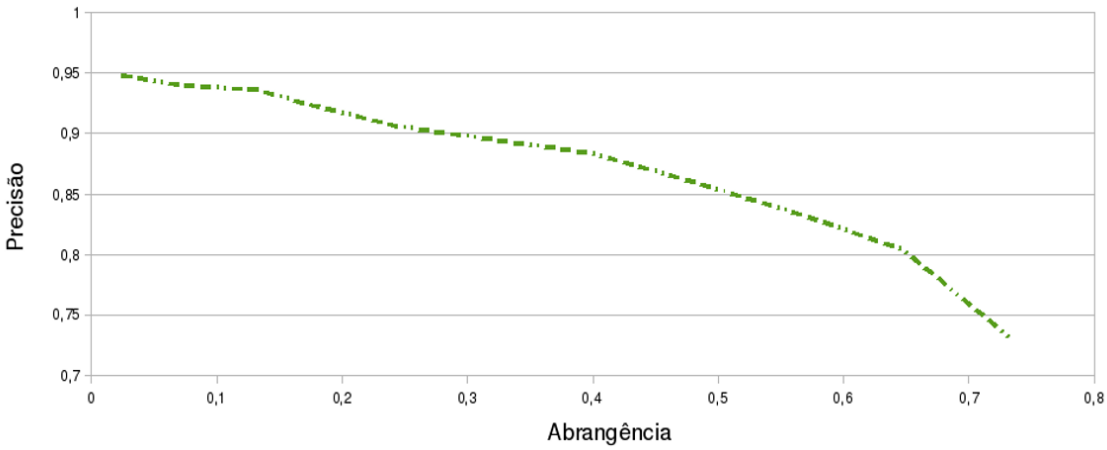

Figura 5.6: Precisão vs Abrangência de Features constituídas por pirâmides de palavras posicionadas consecutivamente à frente e atrás da entidade em causa

Precisão vs Abrangência de Features Constituídas por Palavras Posicionadas Entre Entidades

A curva representada no gráfico da Figura 5.7 apresenta uma Precisão elevada até uma Abrangência de 0.4, depois apresenta um decréscimo acentuado na Precisão fixando-se nos 0.45 para uma Abrangência de 0.45. Estas features obtêm baixa Precisão para valores de Abrangência elevados. De certa forma, estes resultados vêm reforçar a importância da posição das features em relação a entidade, como é possível observar nos gráficos apresentados anteriormente. É possível observar que esta curva assemelha-se muito à curva representada na Figura 5.3. A razão desta semelhança, deve-se ao facto das features serem constituídas por palavras em relação a outras entidades presentes na notícia. Como uma parte significativa das notícias fazem referência a uma só entidade, o classificador ao utilizar somente este tipo de features, não obtém classificação quando lhe é submetido notícias que façam referência a uma única entidade.

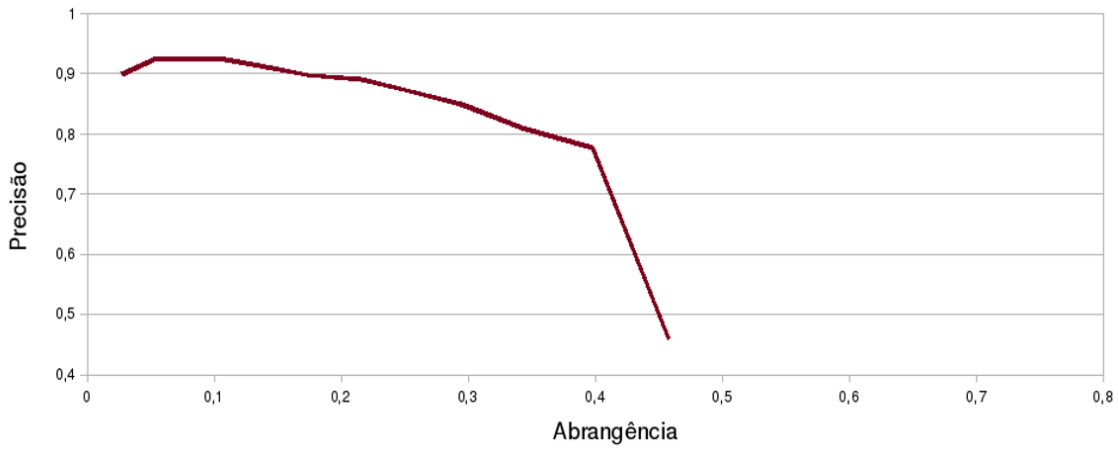

Figura 5.7: Precisão vs Abrangência de Features constituídas por palavras posicionadas entre entidades

Precisão vs Abrangência das Features Constituídas Pelas Características Gramaticais das Palavras

A curva representada no gráfico da Figura 5.8 consegue manter a Precisão de praticamente 0.8 para uma Abrangência entre os 0.1 e 0.4. A partir deste valor de Abrangência apresenta um decréscimo uniforme até uma Abrangência de 0.7. Pelo comportamento desta curva, podemos concluir que a informação gramatical das palavras não é muito importante na representação de features.

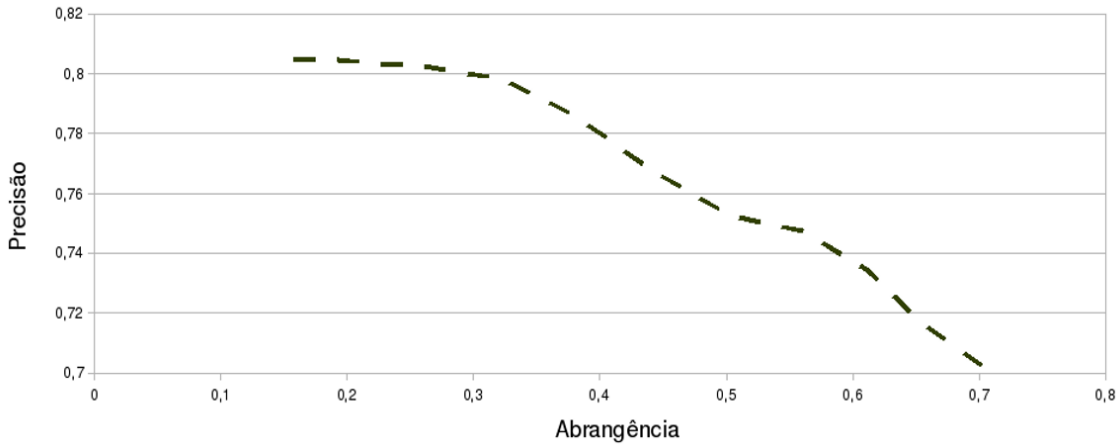

Figura 5.8: Precisão vs Abrangência das Features constituídas pelas características gramaticais das palavras

Precisão vs Abrangência de Todas as Features

O gráfico da Figura 5.9 apresenta o comportamento da Precisão em função da Abrangência utilizando, em simultâneo, todos os tipos de features. É possível verificar que a partir dos 0.2 de Abrangência a curva apresenta um decréscimo praticamente uniforme na Precisão fixando-se aproximadamente nos 0.78, o que corresponde praticamente a uma Precisão de 80%. No entanto, dependendo da aplicação deste classificador e supondo que se pretendia classificar uma amostra de apenas 50% das notícias, que depois se extrapolaria para o total das notícias, estaríamos a conseguir uma Precisão de 88%.

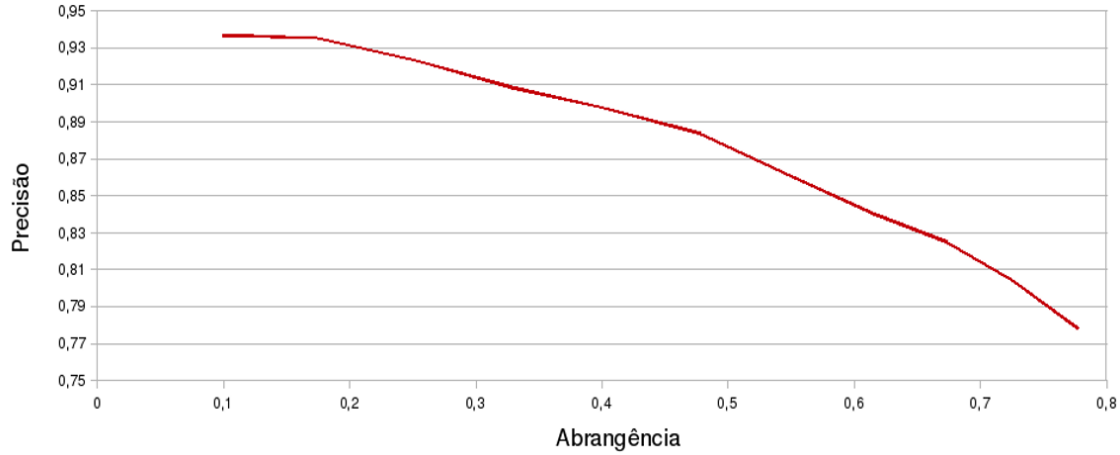

Figura 5.9: Precisão vs Abrangência de todas as Features

5.2 Análise de Resultados do Sistema em Produção

Nesta secção pretende-se analisar os gráficos de valores acumulados e verificar se as tendências que os gráficos apresentam estão, de alguma forma, em consonância com os resultados das Eleições Europeias. Pretende-se avaliar se é viável utilizar este classificador como uma ferramenta que detecte tendências em notícias e dessa forma vir a ser utilizado como uma ferramenta que auxilie à previsão de tendências eleitorais.

Para a realização da análise comparando com resultados reais, serão utilizados os resultados das Eleições Europeias do dia 7 de Junho de 2009 como ponto de controlo.

Análise de Resultados Vital Moreira vs Paulo Rangel

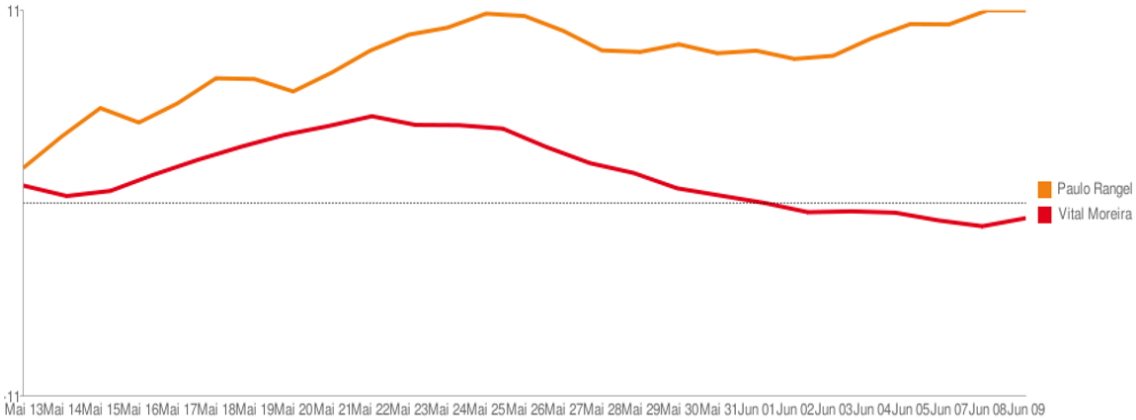

Figura 5.10: Gráfico de tendências acumuladas Vital Moreira vs Paulo Rangel de 13 de Maio a 9 de Junho de 2009

Analisando o gráfico da Figura 5.10 é possível verificar que nos dias 7 e 8 de Junho Vital Moreira atinge o valor mais baixo desde o dia 13 de Maio. No entanto ao contrário do que seria de esperar, Vital Moreira inverte a tendência logo no dia 9 de Junho. A leitura atenta das notícias desse dia permitiu observar que praticamente não existiram notícias a explorar a derrota de Vital Moreira, tendo sido remetida a derrota para o Primeiro Ministro José Sócrates. O tipo de notícias veiculadas pelos meios de comunicação on-line foram do género do exemplo que seguidamente se apresenta e em número muito reduzido:

"Vital Moreira assume responsabilidade pela derrota do PS.
O cabeça de lista do PS, Vital Moreira, assumiu, este domingo, a responsabilidade pela derrota do PS nas eleições europeias, depois de felicitar o PSD."
TSF - 07 de Junho de 2009

Se observarmos o gráfico da Figura 5.11 é possível verificar que, ao contrário da maioria das sondagens que davam vitória a Vital Moreira, já anteriormente a 10 de Maio existia um empate técnico entre os dois candidatos. A partir do dia 10 de Maio, Paulo Rangel destaca-se de Vital Moreira como é possível seguir no gráfico da Figura 5.10.

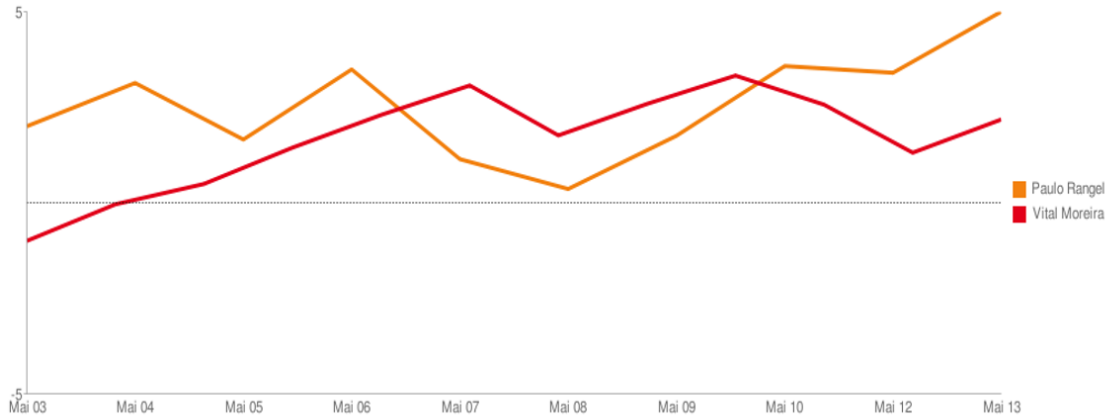

Figura 5.11: Gráfico de tendências acumuladas Vital Moreira vs Paulo Rangel de 3 a 13 de Maio

Análise de Resultados PS vs PSD

Ao analisarmos o gráfico da Figura 5.12 que reflecte a classificação das notícias relativamente às entidades PS e PSD é possível verificar que entre os dias 28 e 31 de Maio, PS e PSD saem de empate técnico que se tinha verificado em períodos anteriores. No entanto, apesar de ser possível verificar uma ligeira subida no dia 8 de Junho para o PSD esta não se prolonga para os dias seguintes. A ausência de uma subida acentuada do PSD, pode ter a sua razão no número de notícias negativas do Bloco de Esquerda e do Partido Comunista Português relativamente ao PS e PSD. No caso do PS estas reflectem-se no gráfico, no caso do PSD estas notícias negativas subtraem-se às positivas, retendo uma subida mais acentuada.

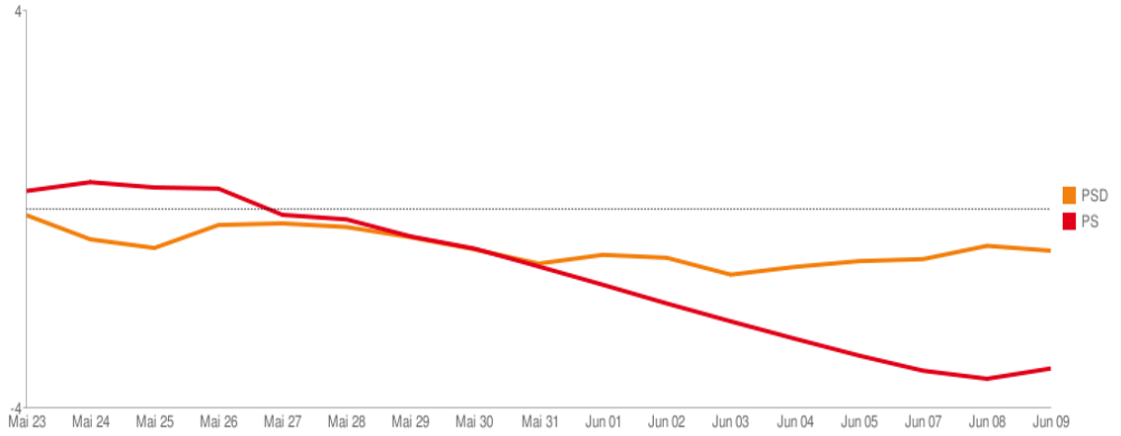

Figura 5.12: Gráfico de tendências acumuladas PS vs PSD de 13 de Maio a 9 de Junho

Análise de Resultados José Sócrates vs Manuela Ferreira Leite

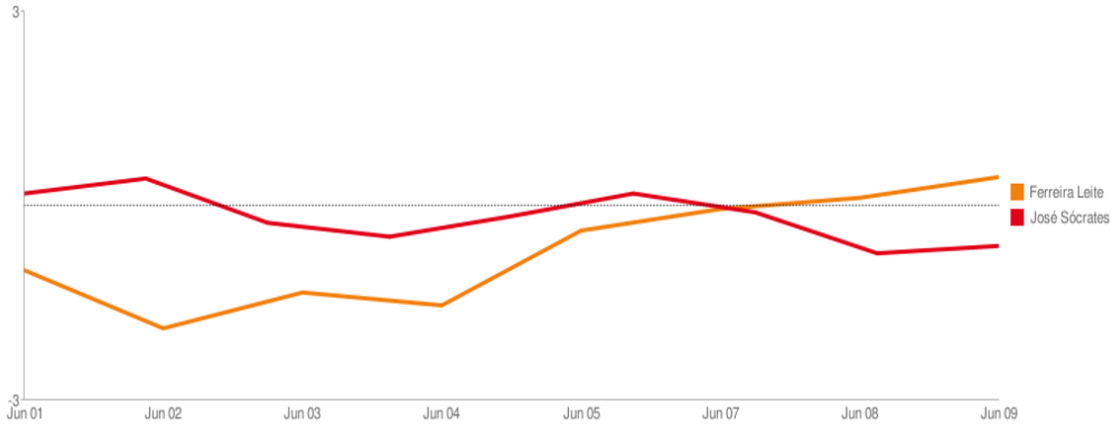

Figura 5.13: Gráfico de tendências acumuladas José Sócrates vs Manuela Ferreira Leite de 1 a 9 de Junho

No gráfico da Figura 5.13 é possível observar que em resultado da vitória do PSD nas Eleições Europeias, o fluxo de notícias favoráveis a Manuela Ferreira Leite e desfavoráveis a José Sócrates, deram origem a que Manuela Ferreira Leite passe de valores negativos para valores positivos e José Sócrates faça o percurso inverso. O gráfico da Figura 5.14 demonstra que a tendência mantem-se até praticamente 23 de Junho, a partir desta data os líderes seguem com empate técnico. Seguidamente são apresentados dois exemplos de notícias sobre o resultado das Eleições Europeias de 7 de Junho de 2009.

"Lisboa, 07 Jun (Lusa) – O cabeça-de-lista social-democrata ao Parlamento Europeu, Paulo Rangel, considerou hoje que a presidente do PSD "é a grande vencedora" destas

eleições europeias e que o secretário-geral do PS sofreu "uma derrota pessoal"."

JN - 07 de Junho de 2009

"Sócrates admite resultado "decepcionante" mas diz que Governo vai manter rumo O secretário-geral do PS, José Sócrates, considerou hoje "decepcionantes" os resultados das eleições europeias, mas frisou que as legislativas serão diferentes e que o Governo vai manter a sua linha de rumo. José Sócrates falava após uma curta declaração do seu cabeça de lista às eleições europeias, Vital Moreira, que assumiu "pessoalmente a derrota"."

PÚBLICO - 07 de Junho de 2009

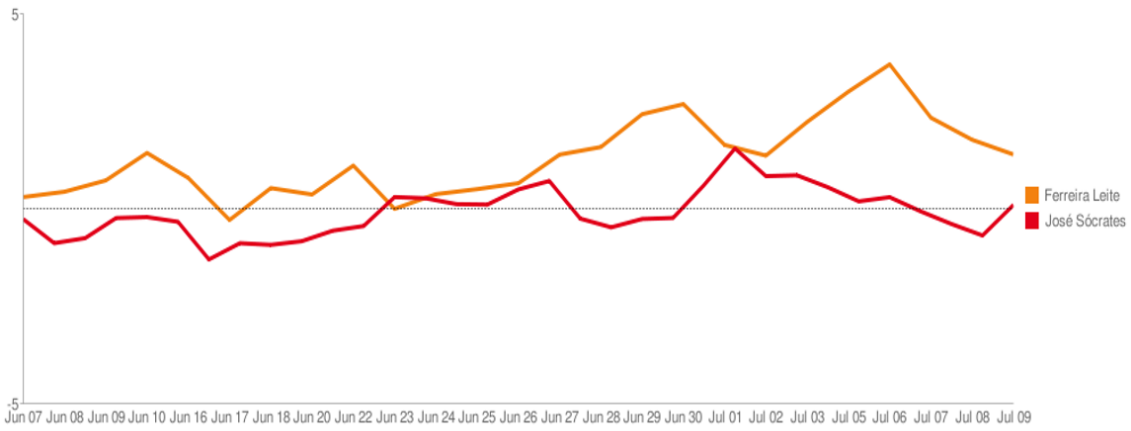

Figura 5.14: Gráfico de tendências acumuladas José Sócrates vs Manuela Ferreira Leite de 7 de Junho a 9 Julho

5.3 Resumo

Neste capítulo procedeu-se à avaliação do classificador segundo as métricas standard. Foram analisados os resultados da utilização dos diversos tipos de features independentemente e em conjunto. Para avaliar os resultados acumulados foram utilizados os resultados das Eleições Europeias, como ponto de controlo. Foi possível verificar que a curva de tendências acompanha os resultados que seriam de esperar, tendo em conta os resultados, nos dias seguintes às eleições Eleições Europeias. Alguns "desvios" que se observaram, relativamente ao que era esperado foram verificados manualmente e concluiu-se que os

"desvios" dos gráficos de tendências eram reflexo da forma como os candidatos transmitiram os resultados das eleições. Importa reter estes resultados uma vez que no Capítulo seguinte (Conclusão) será analisado se o sistema corresponde ao esperado.

Capítulo 6

Conclusões e Trabalho Futuro

Neste capítulo é avaliada a conformidade com os pressupostos inicialmente definidos. São apresentados os resultados e as principais conclusões. Em Trabalhos Futuros são definidos melhoramentos, outros testes e apresentados novos assuntos a explorar.

Conclusões

Podemos concluir que o classificador consegue avaliar a polaridade de notícias de categoria política com uma taxa de Precisão de praticamente 78%. A análise dos gráficos Precisão vs Abrangência, dos diferentes tipos de features, permite concluir que features simples, constituídas por bigramas ou palavras com identificação da posição em relação à entidade, conseguem atingir melhores resultados do que features mais complexas.

Da observação dos gráficos dos valores acumulados das classificações relativamente às Europeias, podemos concluir que o classificador é sensível à polaridade das notícias, uma vez que a tendência dos gráficos no dia 7 de Junho está em concordância com a polaridade das notícias. É possível também verificar que o classificador, ao contrário das sondagens, parece sugerir bastante cedo que Paulo Rangel e o PSD serão os grandes vencedores das Eleições Europeias. Tendo em conta os bons resultados obtidos, pensamos que a análise da polaridade de notícias poderá ser um bom indicador de tendências eleitorais. O classificador automático, apresentado nesta dissertação, consegue realizar esta tarefa eficientemente. No entanto é nossa intenção testar o classificador nas próximas Eleições Legislativas e verificar se são conseguidos bons resultados nessas eleições.

Para além da ferramenta de detecção automática da polaridade de notícias sobre política, outros recursos, especificamente desenvolvidos, resultaram deste trabalho:

Uma base de dados de notícias curtas obtidas de feeds RSS de fontes noticiosas representadas on-line.

Um léxico de entidades políticas, onde se encontram as figuras da política nacional que mais frequentemente surgem na comunicação social.

Uma ferramenta de anotação que agiliza a anotação manual de notícias utilizadas para treinar o classificador.

Trabalho Futuro

Foi possível observar que o classificador baixa significativamente a Precisão em frases com poucas palavras. É nossa intenção, futuramente, analisar e tentar melhorar a Precisão em exemplos com estas características e testar o classificador no Twitter e Facebook. Pretendemos testar combinações das melhores features e outras combinações, de forma a verificar se existem combinações que incrementem a Precisão. Pretendemos também testar a evolução por fonte noticiosa, de forma a detectar "vieses" relativamente aos resultados eleitorais.

Referências

[AG05] Lada A. Adamic e Natalie Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, New York, NY, USA, 2005. ACM Press.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon e Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA, 1992. ACM Press.

[CL08] William W. Cohen e Frank Lin. The multirank? bootstrap algorithm: Semi-supervised political blog classification and ranking using semi- supervised link classification. 2008.

[CV95] Corinna Cortes e Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[CVXS06] Paula Chesley, Bruce Vincent, Li Xu e Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, pages 27–29, Stanford, US, 2006.

[DS06] Kathleen T. Durant e Michael D. Smith. Mining sentiment classification from political web logs. In *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, March 2006.

[FMR98] Johannes Fürnkranz, Tom Mitchell e Ellen Riloff. A case study in using linguistic phrases for text categorization on the www. In *In Working Notes of the AAAI/ICML Workshop on Learning for Text Categorization*, pages 5–12. AAAI Press, 1998.

[GQSE04] Gregory Grefenstette, Yan Qu, James G. Shanahan e David A. Evans. Coupling niche browsers and affect analysis for an opinion mining application. In *Proceeding of RIAO-04*, pages 186–194, Avignon, FR, March 2004.

[GSS07] Namrata Godbole, Manjunath Srinivasaiah e Steven Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[Joa98] Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec e Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Heidelberg, DE, 1998. Springer Verlag.

[KH04] Soo-Min Kim e Eduard Hovy. Determining the sentiment of opinions. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 1367–1373, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[Koh95] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143, San Mateo, CA, 1995.

[Lew95] David D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 246–254, Seattle, Washington, USA, 1995. ACM Press.

[LGDP08] Kevin Lerman, Ari Gilder, Mark Dredze e Fernando Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 473–480, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[MG98] Dunja Mladenic e Marko Grobelnik. Word sequences as features in text-learning. In *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98*, pages 145–148, 1998.

[MKSW99] John Makhoul, Francis Kubala, Richard Schwartz e Ralph Weischedel. Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252, 1999.

[MM06] Tony Mullen e Robert Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, pages 159–162, 2006.

[PLV02] Bo Pang, Lillian Lee e Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Vap95] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.