
Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

# Análise de Sentimento em Documentos Financeiros com Múltiplas Entidades

Javier Zambrano Ferreira

Manaus – Amazonas
Fevereiro de 2014

Javier Zambrano Ferreira

# Análise de Sentimento em Documentos Financeiros com Múltiplas Entidades

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Prof. Marco Antônio Pinheiro de Cristo, Ph.D.

Javier Ferreira

# Análise de Sentimento em Documentos Financeiros com Múltiplas Entidades

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas, como requisito parcial para a obtenção do grau de Mestre em Informática.

Banca Examinadora
Professor Marco Antônio Pinheiro de Cristo, Ph.D. – Orientador
Instituto de Computação – UFAM

Professor Doutor David Braga Fernandes de Oliveira
Instituto de Computação – UFAM

Professor Doutor Álvaro Rodrigues Pereira Júnior
Departamento de Computação – UFOP

Manaus – Amazonas
Março de 2014

*A todos que me acompanharam nessa trajetória.*

# Agradecimentos

Aos meus pais, José Ferreira da Silva e Anita Maria Zambrano Acuña, pois sempre incentivaram meus estudos.

Agradeço também a minha esposa por sempre estar ao meu lado e me incentivando.

Ao meu orientador, meu muito obrigado, pelos direcionamentos, confiança e pelas horas de dedicação.

Agradeço também ao Instituto Nokia de Tecnologia pelo apoio.

*A educação é a arma mais poderosa que você pode usar para mudar o mundo.*
*Nelson Mandela*

# Resumo

Dado o volume de infomação disponível na Internet torna-se inviável a análise manual do conteúdo disponível para identificar diversas informações de interesse. Entre várias análises de interesse, uma de destaque é a análise de polaridade da opinião, ou seja, a classificação de um documento textual em positivo, negativo ou neutro, em relação a um certo tópico. Esta tarefa é particularmente útil no domínio financeiro, onde notícias sobre uma empresa podem afetar o seu desempenho em mercados de ações. Embora a maioria dos métodos nesse domínio considere que os documentos possuem uma única polaridade, observamos que a maioria deles é constituído de múltiplas entidades e o alvo da análise de polaridade é, em geral, as entidades que estes documentos referenciam. O objetivo deste trabalho é, portanto, estudar estratégias para a detecção de polaridade em documentos financeiros com múltiplas entidades. Para tanto, estudamos métodos baseados na criação de múltiplos modelos de aprendizado com um conjunto pré-definido de entidades, usando o classificador SVM. Nós avaliamos tanto modelos baseados em conjuntos de documentos específicos por entidade quanto modelos baseados em segmentação de documentos usando diversas heurísticas de processamento de linguagem natural. Os resultados mostraram que há um ganho em fragmentar os textos para análise de polaridade com rótulos de classificação por entidades.

**Palavras-chave:** Análise de Polaridade, Múltiplas Entidades, SVM, Processamento de Linguagem Natural, Resolução de Anáforas

# Abstract

Given the amount of information available on the internet, it becomes unfeasible the manual content analysis to identify information of interest. Among such analyses, one of particular interest is the polarity analysis, that is, the classification of a text document in positive, negative, and neutral, according to a certain topic. This task is particularly useful in the finance domain, where news about a company can affect the performance of its stocks. Although most of the methods about this domain consider that documents have only one polarity, in fact most of the documents cite many entities and these entities are often the target of the polarity analysis. Thus, in this work, we intend to study strategies for polarity detection in financial documents with multiple entities. In particular, we study methods based on the learning of multiple models, one for each observed entity, using SVM classifiers. We evaluate models based on the partition of documents according to the entities they cite and on the segmentation of documents into fragments according to the entities they cite. To segment documents we use several heuristics based on shallow and deep natural language proecssing. We found that entity-specific models created by simply partitioning the document collection largely outperformed strategies based on single models.

**Keywords:** Polarity Analysis, Multiple Entities, Machine Learning, Natural Language Processing, Anaphora Resolution

# Conteúdo

# Lista de Figuras

# Capítulo 1

# Introdução

Dado o volume de informação disponível na Internet, seja em *Websites*, fóruns e página de notícias, é inviável a análise manual do conteúdo disponível para identificar diversas informações de interesse. Um tipo de análise de interesse desse conteúdo consiste em determinar a polaridade da opinião do autor em relação ao assunto em discussão, o que chamamos de análise de sentimento ou polaridade. Um exemplo de análise de sentimento é inferir se, em um texto sobre um produto, o autor do texto emite uma opinião favorável, neutra ou desfavorável em relação ao produto.

A análise de sentimentos tem sido usada em uma varidade de domínios de aplicação. Por exemplo, ela é útil para inferir automaticamente a opinião de um cliente sobre um certo produto de uma loja virtual com base em um comentário que este postou, a opinião de uma pessoa sobre um item postado em uma rede social, etc. Enquanto algumas técnicas gerais podem ser usadas para qualquer domínio, a simples transposição do que vale em um domínio para o outro pode não ser bem sucedida. Como observado por [18], diversos termos pré-classificados como positivos em um domínio possuem uma conotação neutra em diferentes contextos.

Um domínio de particular interesse, e foco deste trabalho, é o domínio dos documentos financeiros. O interesse neste domínio se deve à hipótese de que notícias de caráter positivo ou negativo, relacionadas com uma companhia, podem afetar o desempenho financeiro desta companhia na bolsas de valores [1, 8, 9]. Assim, a polaridade de um documento de natureza financeira poderia ser usada para ajudar

a prever tendências relacionadas com o desempenho de uma companhia.

Em termos de desenvolvimento de técnicas e algoritmos, o desafio, no caso de documentos financeiros, é maior, uma vez que, ao contrário de domínios como filmes e produtos, os autores dos textos não avaliam as entidades citadas por meio de avaliação do conteúdo com notas que podem variar de 0 a 5, por exemplo. [1]. Outra característica que observamos nos trabalhos realizados neste domínio é que estes parecem assumir que os documentos tem uma única polaridade, independente das entidades citadas [1, 16]. Neste trabalho, não assumimos esta premissa, uma vez que diversos documentos citam duas ou mais entidades e, muitas vezes, com diferentes polaridades. Como um simples exemplo, citamos o texto dado na Figura 1.1.

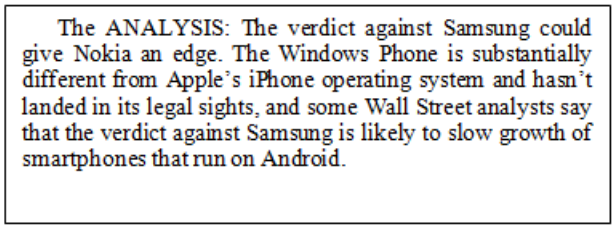


Figura 1.1: Exemplo de texto com múltiplas entidades.

Como podemos observar na Figura 1.1, três entidades são citadas no texto: *Nokia*, *Apple* e *Samsung*. Note nesse exemplo, contudo, que a polaridade do texto é diferente para cada entidade, uma vez que ele informa que a Nokia pode ganhar com a perda da Samsung em um julgamento que envolveu Samsung e Apple. Em particular, este documento é neutro em relação à Apple uma vez que apenas a cita como fabricante do iPhone, ao mesmo tempo que é positivo em relação à Nokia e negativo em relação à Samsung.

Até onde sabemos, a abordagem sugerida em trabalhos anteriores (como em [1, 8]), iria inferir a polaridade do texto como um todo para as entidade citadas. Ou seja, se o algoritmo de análise de sentimentos considerasse o texto dado no exemplo anterior como negativo, esta seria a polaridade atribuida à Apple, Samsung e Nokia, o que estaria claramente incorreto.

| Número de Entidades | Total de Documentos | Percentual |
|---|---|---|
| 1 | 9 | 0% |
| 2 | 214 | 6% |
| 3 | 483 | 13% |
| 4 | 650 | 17% |
| 5 | 730 | 19% |
| 6 | 571 | 15% |
| 7 | 427 | 11% |
| 8 | 268 | 7% |
| 9 | 182 | 5% |
| 10 | 120 | 3% |
| 11 | 84 | 2% |
| 12 | 33 | 1% |
| 13 | 18 | 0% |
| 14 | 10 | 0% |
| $\geq 15$ | 32 | 1% |

Tabela 1.1: Documentos com o número de entidades citadas em seu texto.

De fato, a existência de documentos financeiros que citam múltiplas entidades é relativamente comum. Podemos atestar isto em uma coleção de 3821 páginas catalogadas da seção financeira da *Bloomberg* que coletamos em 2012. A Tabela 1.1 mostra o total de documentos com o número de entidades citadas no texto. As entidades consideradas são as 2000 maiores companhias constanstes da lista *Forbes 2000*[1].

Nesta tabela, observamos que 80% dos documentos citam entre duas a oito entidades, com o número mais comum de entidades por documentos igual a cinco. Um exemplo de documento nesta coleção é dado na Tabela 1.2, que descreve as ações financeiras de várias entidades, a saber, *Google*, *IBM*, *Intuitive Surgical Inc.*, *JDS Uniphase Corp.*, *Marathon Petroleum* e *Microsoft*. O texto é sobre a queda nas ações do Google e a alta nas ações da Microsoft, além de descrever outras entidades de ramos distintos. Neste caso, em particular, observamos que o documento é negativo em relação ao Google e positivo em relação à Microsoft.

Nesta mesma coleção, também podemos observar que muitos dos documentos que

[1]http://en.wikipedia.org/wiki/Forbes_Global_2000

| |
|---|
| Google Inc. (GOOG) fell 8.4 percent to $585.99, after losing 8.4 percent for the biggest loss in the Standard & Poor's 500 Index. The owner of the world's most popular Internet search engine reported fourth-quarter revenue and profit that missed analysts' estimates as an economic slowdown in Europe crimped international sales.<br>Insmed Inc. (INSM) surged 32 percent, the biggest gain in the Russell 2000 Index, to $5.01. The Food and Drug Administration lifted a suspension of clinical trials of the company's experimental drug Arikace for patients with non- tuberculous mycobacteria lung disease.<br>International Business Machines Corp. (IBM) advanced 4.4 percent, the most since July 19, to $188.52. The world's biggest computer-services provider forecast 2012 earnings exceeding analysts' estimates after fourth-quarter profit rose 4.4 percent because of rising software demand.<br>Intuitive Surgical Inc. (ISRG) sank 6.1 percent, the most since Aug. 8, to $445.68. The maker of a robotic system to perform surgery said annual growth in its da Vinci surgical procedures slowed to 27 percent in the fourth quarter from 30 percent in the third quarter.<br>JDS Uniphase Corp. (JDSU) rose 4 percent to $13.45, the highest price since Sept. 15. The maker of fiber-optic equipment was raised to "buy" from "hold" at Stifel Nicolaus & Co., which said the company has a "sustainable competitive advantage."<br>Marathon Petroleum Corp. (MPC) (MPC US) jumped 3.7 percent to $37.17, the highest price since Nov. 11. The largest independent U.S. refiner by market value outperformed competitors after hedge fund Jana Partners LLC bought a 5.5 percent stake in the company.<br>Microsoft Corp. (MSFT) climbed 5.7 percent, the biggest gain in the Dow Jones Industrial Average, to $29.71. The world's largest software maker reported second-quarter profit that beat estimates, lifted by holiday sales of Xbox machines and Kinect sensors, as well as corporate software demand. |

Tabela 1.2: Exemplo de documento com descrição das ações no mercado financeiro.

citam mais de uma entidade, possuem polaridades distintas para estas entidades. Por exemplo, a Tabela 1.3 apresenta as polaridades destes documentos em relação à Apple quando comparada com outras quatro entidades (Microsoft, Google, Nokia e Samsung) em uma amostra de cerca de cinquenta documentos.

Do conjunto de 3821 documentos examinados, notamos que 38% das polaridades atribuídas são idênticas entre a Apple e outra companhia (32% positivas e 6% negativas). Das atribuições restantes (62%) com polaridades distintas, 38% são positivas pra Apple e negativas pra outra entidade. Finalmente, 23% apresentam polaridades negativas pra Apple e positivas pra outra entidade. Estes resultados confirmam que diferentes polaridades podem ser observadas em um mesmo documento.

| | | Microsoft | | Google | | Nokia | | Samsung | |
|---|---|---|---|---|---|---|---|---|---|
| | | POS | NEG | POS | NEG | POS | NEG | POS | NEG |
| Apple | POS | 8 | 12 | 3 | 0 | 6 | 11 | 4 | 2 |
| | NEG | 6 | 1 | 4 | 2 | 1 | 0 | 4 | 1 |

Tabela 1.3: POS: positivo e NEG: Negativo. Entidade Apple com relação as demais.

A observação de documentos com múltiplas entidades com polaridades distintas claramente motiva o estudo de estratégias de análise de sentimentos multi-entidade. Em um contexto de aprendizagem de máquina, o problema pode ser pensado como uma tarefa de multi-classificação em que cada documento tem uma polaridade para cada entidade de interesse. Em outras palavras, em lugar de se treinar um único modelo por documento (analisando os rótulos positivo, negativo ou neutro para o documento como um todo), se buscaria aprender um modelo por entidade (analisando rótulos positivo, negativo ou neutro para cada entidade de interesse $E_1, E_2, ..., E_n$).

Uma primeira abordagem de aprendizagem seria a criação dos modelos de aprendizado específicos considerando toda a coleção. Uma segunda seria a criação de modelos a partir de bases em que todos os documentos citam a entidade-alvo. A hipótese por trás desta ideia é que o conjunto de documentos relacionados com uma entidade específica fornece material mais adequado para o aprendizado de polaridade para aquela entidade. Estendendo esta ideia para documentos individuais, uma nova hipótese seria que fragmentos de um documento que se relacionam a uma entidade formam uma base melhor para o aprendizado de polaridade. Logo, a base para o aprendizado do modelo por entidade seria o conjunto de fragmentos de cada documento que se referem diretamente à entidade alvo.

Neste trabalho, partindo dessas ideias, propomos diversas estratégias para o aprendizado de modelos multi-entidades para a tarefa de análise de polaridade. Avaliamos os vários modelos propostos comparando-os com a ideia tradicional de se considerar apenas uma polaridade por documento. Como resultado de nossa

avaliação, como esperado, observamos que todas as ideias baseadas em múltiplas entidades apresentaram melhor desempenho que a estratégia baseada em única polaridade. Entre as estratégias baseadas em múltiplos modelos (por entidade), a melhor foi a que segmenta os documentos usando uma técnica de processamento de linguagem natural para a resolução de anáforas (cf. Seção 3). Contudo, o seu desempenho foi pouco superior a técnicas muito mais simples baseadas em segmentação do texto de acordo com a observação da entidade alvo.

## 1.1 Objetivos

### 1.1.1 Geral

Desenvolver um método eficaz para identificar polaridade associada às entidades em documentos textuais no domínio financeiro. As polaridades a serem detectadas serão referentes às entidades citadas no documento e as estratégias estudadas são baseadas em modelos distintos para entidades distintas.

### 1.1.2 Específicos

Este trabalho possui os seguintes objetivos específicos:

1. Criar uma coleção de documentos financeiros para análise de polaridade, com múltiplas entidades. Para esta coleção, verificar a distribução das entidades e, para uma amostra dela, rotulá-las de acordo com as polaridades observadas, tendo usado avaliadores humanos;

2. Implementar método da literatura que será usado como base de comparação;

3. Verificar como diferentes estratégias para o aprendizado de múltiplos modelos se comparam ao método de base. As estratégias a serem estudadas foram propostas ao longo da pesquisa e são brevemente descritas na Seção 1.2

## 1.2 Metodologia

O trabalho foi implementado de acordo com as etapas descritas nos parágrafos a seguir.

A primeira etapa consistiu em uma pesquisa bibliográfica envolvendo análise de polaridade em diversos domínios, tendo como prioridade os trabalhos no domínio financeiro. Com base nesta fase, determinou-se a estratégia geral para o aprendizado de polaridade de texto, baseado no estado-da-arte da literatura.

A segunda etapa consistiu em implementar o algoritmo de análise de sentimentos selecionado, ou seja, o método proposto em [1]. O algoritmo consiste na extração da terminologia dos documentos, por meio de métodos de processamento de linguagen natural como reconhecimento de parte-do-discurso, identificação de variação no uso de termos e o uso de métodos de ponderação baseados nas freqüências dos termos (TF e IDF). Os termos relevantes no domínio financeiro foram então aprendidos por meio de um classificador Support Vector Machine [19, 2].

A terceira etapa foi a criação de uma coleção para os experimentos. Inicialmente, foram coletadas páginas financeiras, em língua inglesa, publicadas nos *sites* *Bloomberg*, *New York Times*, *Financial Times*, *Reuters* e *AllthingsD*. Considerando que estes documentos deveriam ser rotulados para múltiplas entidades, manualmente, optou-se por definir cinco entidades de interesse para viabilizar a rotulação. As entidades escolhidas foram Apple, Google, Nokia, Microsoft e Samsung. Das páginas coletadas originalmente, foram então extraídas mil que citavam uma ou mais destas cinco entidades. Estes documentos foram avaliados por anotadores humanos que os rotularam como positivo, negativo ou neutro para cada uma das entidades. Também foram obtidas avaliações globais para cada um dos documentos.

A quarta etapa consistiu na criação de diferentes estratégias para o aprendizado de modelos específicos por entidades. Para a identificação de referências a uma entidade por meio de resolução de anáforas, usamos a ferramenta que representa o estado da arte nesta tarefa: *Beautiful Anaphora Resolution Toolkit* (BART) [17].

Descrições detalhadas dessas estratégias são fornecidas no Capítulo 3.

A última etapa do nosso trabalho é a avaliação dos resultados. Para tanto, com base na coleção criada manualmente, iremos avaliar a acurácia dos diversos métodos estudados.

## 1.3 Organização da Dissertação

Esta dissertação está dividida da seguinte maneira: o Capítulo 2 apresenta os conceitos gerais para o entendimento do trabalho realizado, incluindo uma descrição dos trabalhos relacionados ao nosso; o Capítulo 3 descreve a nossa proposta para análise de entidades considerando ou não segmentação dos documentos de acordo com as entidades citadas. Os experimentos e os resultados obtidos são descritos no Capítulo 4 e, por fim, no Capítulo 5, apresentamos nossas conclusões.

# Capítulo 2

# Conceitos Básicos

Neste capítulo apresentaremos conceitos gerais usados neste trabalho para técnicas de análise de sentimentos, anafóras e métricas de avaliação.

## 2.1 Análise de Sentimentos

A Análise de Sentimentos é uma subárea de Aprendizagem de Máquina e Processamento de Linguagem Natural, o qual tem o objetivo geral de inferir o sentimento expresso pelo autor do texto em relação a um certo objeto (por exemplo, o tópico do texto, um produto, serviço ou pessoa citado no texto) a partir do seu conteúdo. Dentre as análises realizadas, uma muito comum é a Análise de Polaridade que consiste em determinar se o texto descrito é positivo, negativo ou neutro em relação ao seu objeto. A análise de polaridade tem inúmeras aplicações, entre as quais citamos:

- Reviews: *Sites* de comércio eletrônico, filmes, música e outros, permitem que usuários avaliem seus conteúdos por meio de notas. Entretanto, há casos em que usuários comentam de forma positiva, porém a nota de avaliação é baixa. Com o uso de análise de polaridade é possível verificar e corrigir tais distorções.

- Componente: o uso de análise de sentimento como componente em sistemas de recomendações ajuda em não recomendar produtos com *reviews* negativos;

Em sistemas de propagandas, quando o texto tiver contexto positivo mostra-se o produto e o mais importante, em contexto negativo o produto não aparece.

- Produtos: Um produto pode ser avaliado como um todo ou em partes. Por exemplo, um celular pode ter avaliação positiva, porém sua camêra é avaliada como negativa.

- Mercado Financeiro: uma das principais característica do mercado financeiro é a avaliação dos investidores e profissionais sobre uma companhia. Essas opiniões podem afetar o mercado de ações e influenciar para mais ou para menos o preço da ação. Por meio da análise de sentimentos é possível detectar opinões positivas e negativas, o que ajuda os investidores a tomarem decisões.

Uma técnica simples para a implementação de aplicações de análise de sentimento é a baseada na análise de palavras individuais (*bag of words*). Nesta técnica, cada palavra em um documento é classificada individualmente quanto à sua polaridade: positiva, negativa ou neutra. A avaliação do termo como membro de uma classe (positiva, negativa ou neutra) é por meio de um dicionário com informação de polaridade como, por exemplo, no caso do Inglês, a WordNet[1]. De acordo com a Wordnet, palavras como *good* e *happy* tem polaridade positiva, enquanto palavras como *bad* e *sad* tem polaridade negativa. A polaridade de um documento é a mais comum entre os seus termos. Ou seja, se os termos mais freqüentes são em sua maioria positivos, o documento é positivo [13]. Como a frequência dos termos dependem do seu domínio e o significado depende do contexto em que são usados, o uso de um conjunto de termos de um domínio em outro pode diminuir a acurácia do classificador de polaridades.

Outros conjuntos de técnicas se baseiam em informações lingüísticas mais complexas como a função do termo no discurso (*part-of-speech*) ou a posição que o mesmo aparece no texto. Estes métodos se baseiam no uso de heurísticas como,

[1]WordNet: http://wordnet.princeton.edu/

por exemplo, dado o unigrama BOM, rotulado como adjetivo (ADJ), o seu peso é menor no início de uma sentença (ex: "Bom, apesar do produto ter um ótimo desempenho...") que no fim (ex: "O produto é muito bom"). No primeiro caso, nota-se que o termo BOM foi utilizado como introdução conclusiva, enquanto no segundo, foi usado como um adjetivo positivo. Além disso, é possível identificar relações entre os termos e entender o significado de um conjunto de termos em sequência (n-gramas) que identificam uma classe de polaridade. Assim, o n-grama "muito bom" pode ser interpretado como mais positivo que "bom" enquanto o n-grama "não é muito bom" é compreendido de forma negativa dada a inversão de sentido do termo "não". Diversas técnicas utilizam processamento de linguagem natural, como exemplificas em [13, 12]. O trabalho dos autores em [12] também analisa várias estratégias complexas para a interpretação de sentenças.

Além das técnicas que visam derivar polaridades de palavras ou sentenças individuais por meio de heurísticas mais ou menos complexas baseadas em processamento de linguagem natural, a análise de sentimento também pode ser vista como um problema de classificação. Neste caso, muitas das informações usadas em métodos anteriores, bem como unigramas ou n-gramas, podem ser usados diretamente como atributos associados aos documentos que se pretende classificar. Assim, classificadores são utilizados em diversos trabalhos [13]. Vários métodos tem sido usados na área, entre os quais, baseados em classificadores supervisionados, não supervisionados e semi-supervisionados. Os classificadores supervisionados requerem uma fase de aprendizagem tendo como entrada um conjunto de treinamento. Tal conjunto de treinamento contém documentos rotulado por humanos, por isso o termo supervisionado. Já no classificador não supervisionado, a fase de aprendizado não existe. Exemplos comuns desses métodos são os baseados em agrupamento. Por fim, no caso dos semi-supervisionados utiliza-se um pequeno conjunto de base de treino rotulado e outro não rotulado [2, 19].

Para formalizar o problema de classificação em análise de sentimentos, tomamos como base a definição em [2] para classificação de texto: "Dado um conjunto $\mathcal{D}$ de documentos e um conjunto $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ de $n$ classes com seus respectivos rótulos, um classificador é uma função binária $\mathcal{F} : \mathcal{D} \times \mathcal{C} \implies \{0, 1\}$.". Tal função binária é responsável em definir 0 ou 1 para um conjunto de par $[d_j, c_i]$ tal que $d_j \in \{D\}$ e $c_i \in \{C\}$. Caso o documento $d_j$ pertença a classe $c_i$, $\mathcal{F}(d_j, c_i)$ é 1, sendo 0 no caso contrário. No contexto de análise de sentimentos as classes $\mathcal{C}$ poderiam ser interpretadas como diferentes polaridades (negativo, positivo ou neutro). Note que a classificação pode ser feita para um único rótulo ou para múltiplos rótulos. O problema para a classificação de um único rótulo é mais díficil, pois transcende apenas a decisão de ser ou não da classe. Isso porque, além de classificar, precisa-se definir qual é a melhor classe a ser definida para o novo documento. Em classificadores com multiplos rótulos, um documento pode ser classificado para mais de um tipo de classe ou único documento.

Neste trabalho utilizamos um classificador supervisionado. Por ser supervisionado, um grupo de pessoas avaliaram 1000 documentos em três possíveis classes: positivos, negativos e neutros, isso de acordo a cada entidade citada: Apple, Google, Microsoft, Nokia e Samsung. Em seguida, separou-se os documentos em um conjunto de teste e outro conjunto de treino. A Figura 2.1 apresenta um fluxograma das etapas de um classificador supervisionado[2].

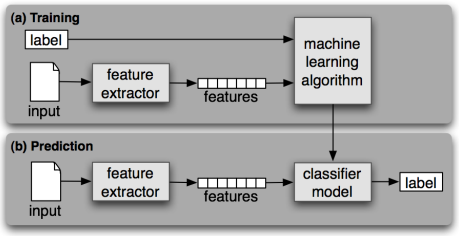


Figura 2.1: Representação das etapas de um classificador supervisionado.

[2]http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html.

Como observado na Figura 2.1, na fase de treinamento, dada a coleção de treino, são extraídos as atributos de interesse. Em nosso caso, unigramas após vários tratamentos linguísticos como eliminação de palavras funcionais (stopwrods) e redução ao radical (stemming). Estes termos são entregues então ao classificador que cria um modelo dos padrões dos atributos associados a cada classe de polaridade. Na fase de teste, dado um novo documento, seus atributos são extraídos e fornecidos como entrada ao modelo criado no treino para que seja possível definir a polaridade do documento.

Entre os diversos tipos de classificadores supervisionados [2, 1, 13], utilizamos o *Support Vector Machine* (SVM) [19, 2]. O SVM é um método de classificação binária que utiliza-se da teoria de vetores e espaços vetoriais para obter a classificação de objetos que podem ser representados por vetores. No SVM, os termos de um documento são representados como pontos ou vetores no espaço. Com a representação dos documentos em vetores, a ideia do SVM é encontrar o hiperplano ótimo que melhor divide as duas classes $C_a$ e $C_b$. Esse hiperplano é determinado por meio do conjunto de treino com os documentos já classificados por seres humanos. Na Figura 2.2 tomamos como exemplo que os círculos pertencem a classe $C_a$ e os pontos representados pelo quadrado são os documentos da outra classe, $C_b$. Já o vetor $w$ é o vetor gerador do hiperplano ótimo e está representado pela linha verde. Porém, há casos que não é possível que haja um hiperplano que divida linearmente as duas classes. Portanto, deve-se considerar que possam ocorrer alguns erros. Para isso, uma margem de erros aceitáveis é definida, por meio de variáveis de folga. A Figura 2.3 mostra um documento $C_b$ classificado erradamente.

Para a classificação de novos documentos representa-se o mesmo por pontos no plano e verifica-se em que lado do hiperplano ele se encontra. Quanto mais próximo do vetor $w$ apresentados na Figura 2.2 e 2.3, mas difícil classificá-lo. Portanto, documentos que estão mais distantes do vetor $w$, pode-se dizer que estão classificados com uma maior certeza.

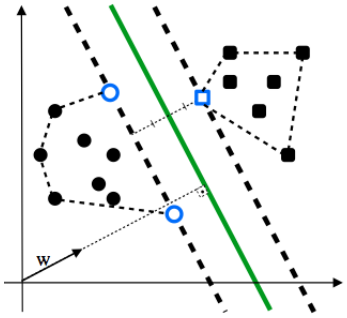


Figura 2.2: Os pontos representados por círculos pertencem a classe $C_a$ e os pontos representados por quadrados a classe $C_b$. O vetor $w$ é o vetor gerador do hiperplano ótimo, representado pela linha verde.

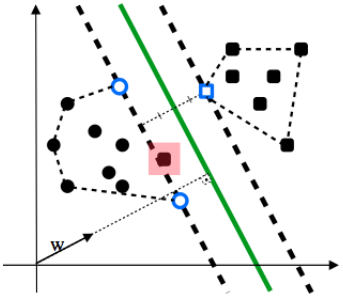


Figura 2.3: Representação de um ponto da classe $C_b$ do lado do hiperplano da classe $C_a$.

Em diversos casos, o SVM é utilizado para classificar múltiplas classes, mesmo tendo sua origem como classificador binário. Para isso, usa-se a estratégia de reduzir um problema de múltiplas classes para um grande problema de classificação binária. Uma maneira simples, apresentado em [2], é considerar um problema binário por classe. Para cada classe, o classificador verificava se o documento pertence a uma classe particular. Ao fim, as várias decisões são combinadas para se ter a decisão final.

Para a utilização das técnicas citadas, duas ferramentas foram usadas: Weka[3] [6] e LIBSVM[4] [5]. A ferramenta Weka consiste em um arcabouço de algoritimos de aprendizagem de máquina para questões de mineração de dados por meio da linguagem Java. Para este trabalho, utilizamos a ferramenta Weka para a filtragem de

[3]http://www.cs.waikato.ac.nz/ml/weka/
[4]http://www.csie.ntu.edu.tw/c̃jlin/libsvm/

dados, criação de partições de treino e teste e interface para o classificador LIBSVM. O LBSVM foi usado para a criação dos modelos para cada classe de polaridade. Os dados foram normalizados para garantir média zero e variância unitária. O método SVM usado foi o linear com escolha de parâmetros (nesse caso, o custo associado com erros ocorrendo na folga da margem) feita pela ferramenta de busca sequencial *grid*, fornecida com o próprio LIBSVM.

## 2.2 Resolução de Anáforas

Na seção 3, apresentaremos as proposta para segmentação de texto de acordo com a entidade referenciada em cada parágrafo. Uma abordagem trivial para determinar se uma certa entidade ocorre em um texto é feita por simples casamento do nome da entidade. Por exemplo, dadas as sentenças *João realmente gosta de chocolate. Contudo, ele sempre suja o seu rosto quando come isso*, é possível inferir que a primeira sentença se refere à entidade João, uma vez que o nome João ocorre nela. Contudo, não é tão simples concluir o mesmo para a segunda sentença, onde a entidade João foi representada pelo pronome "ele".

O problema de determinar que partes de um texto referenciam uma mesma entidade é chamada de resolução de anáforas ou resolução de co-referências. Este problema surge do fato de que uma entidade pode ser referida através de diferentes expressões linguísticas. Nas sentenças dadas anteriormente como exemplos, nota-se que *João, ele e seu* referenciam a entidade *João* enquanto *chocolate* e *isso*, a entidade *chocolate*.

Uma definição formal para o problema de anafóra é: um substantivo A é dito um antecedente anafórico de B se e somente se A é necessário para a interpretação de B. Portanto, anafóra é irreflexiva, assimétrica e não transitiva.

A solução de anafóra tem sido tópico de diversas pesquisas em Processamento de Linguagem Naturais (PLN) [15, 10]. A identificação de substantivos em uma

frase é essencial para a solução de anáforas, isso porque substantivos geralmente descrevem pessoas, lugares, coisas ou conceitos. Portanto, uma entidade pode ser descrita como um substantivo em destaque na frase [15, 10] e para identificar tais substantivos é necessário um conhecimento da gramática da linguagem estudada. No caso da lingua inglesa, usada neste trabalho, usa-se os seguintes conceitos para encontrar um substantivo:

- Pronome definitivo: tal como *he* ou *she*, *Ross bought* {*a radiometer / three kilograms of after-dinner mints*} *and gave* {*it / them*} *to Nadia for her birthday.*
- Pronomes indefinidos como *one* em *Kim bought a t-shirt so Robin decided to buy one as well.*
- Pronomes Demonstrativos tal como *that.*
- Nominais: *a man*, *a woman* e *the man.*
- Nomes próprios: John, Mary, *Amazon*, etc.

Além disso, é importante identificar quais palavras de um texto são verbos, adjetivos, preposições e advérbios visto que as regras gramáticas de uma linguagem definem a ordem da apresentação do substantivo na frase. Com o conhecimento das regras da linguagem estudada, algoritmos de resolução de anafóras [15, 10] podem fazer uso de técnicas de *part-of-speech tagging.* Tais técnicas conseguem identificar a categoria lingüística que um *token* pertence, a qual é definida pelo comportamento sintático ou morfológico do item. Outro aspecto importante para definir a classe que pertecem é o contexto em que o termo é utilizado, pois não é raro um termo ser ambíguo.

Em geral, técnicas de *part-of-speech tagging* recebem como entrada um texto, que é dividido em *tokens* que, por meio de um corpo lingüístico, são classificados conforme a classe a que pertencem. Em destaque, dois corpos são utilizados em

diversos trabalhos da literatura [10, 15, 3]: Brown Corpus[5] e WordNet[6]. A Figura 2.4 exemplifica a saída de um algoritmo de *part-of-speech*.

```
>>> text = nltk.word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

Figura 2.4: Exemplo da saída de um algoritmo de part-of-speech.

Um dos primeiros algoritmos com abordagem linguística que utiliza-se de árvores para fazer um parse sintático de sentenças é o *Hobb's* [10, 15, 3]. Nota-se que na Figura 2.5 há uma árvore para a sentença: *John likes him*. Com a identificação que John é um nome próprio, *NP = Proper noun* e *him* também. O algoritmo navega em largura na árvore da esquerda para a direita e identifica que quanto maior a profundidade da folha direita em que o termo é classificado como NP, maior a chance de referenciar o termo mais alto do lado esquerdo da árvore.

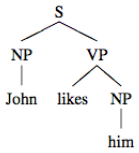


Figura 2.5: Exemplo de uma árvore de parse sintática do algoritmo de Hobb's.

Outras técnicas utilizam o aprendizado de máquinas por meio de uma base de conhecimento. Diversos corpos linguísticos são disponibilizados na literatura [3]. Por meio desses corpos é possível utilizar classificadores supervisionados, isso porque como houve uma classificação de termos por seres humanos a base de treino está criada. Um tipo de classificador supervisionado utilizado são as árvores de decisão [15, 10]. A Figura 2.5 do algoritmo de Hobb's lembra tal árvore. Entretanto, em árvores de decisão, cada nó da árvore toma-se uma decisão com relação ao rótulo ali contido [10]. Por exemplo, a distância entre o nó com maior profundidade do lado esquerdo (folha) com relação a raiz determina que o termo contido na folha é uma entidade e que o termo na raiz é referente a ele.

[5]http://en.wikipedia.org/wiki/Brown_Corpus
[6]http://wordnet.princeton.edu/

Neste trabalho, utilizamos a ferramenta *BART* (Beautiful Anaphora Resolution Toolkit)[7] que representa o estado-da-arte para a resolução de anáforas [17]. Inicialmente BART foi criado para a linguagem inglesa, porém é possível utilizá-lo para diversas linguas como alemão, italiano e outros.

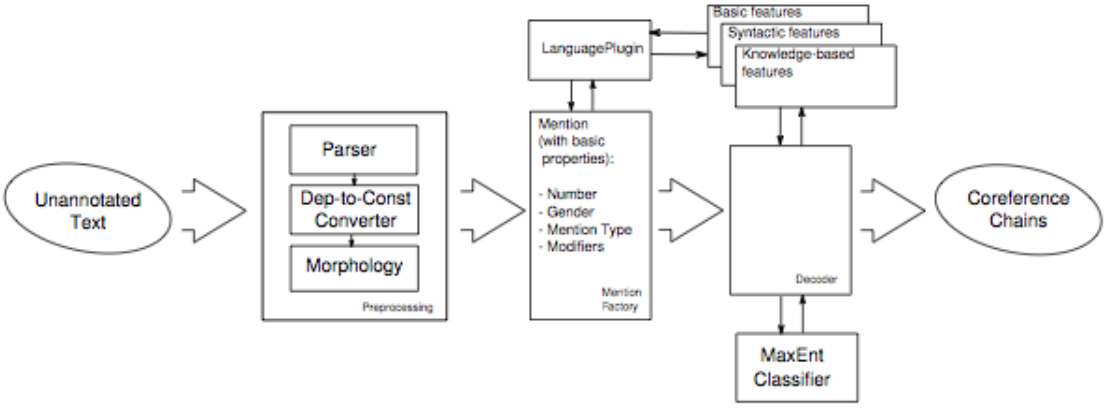


Figura 2.6: Arquitetura da ferramenta BART.

A ferramenta possui cinco principais componentes: um *pipeline* de pré-processamento, detecção de menções, extração de características, decodificador e codificador. Além disso, ela possui um módulo de linguagem com informações específicas da linguagem a ser analisada. A Figura 2.6 representa a arquitetura dos componentes citados. Cada módulo pode ser acessado de forma independente, o que permite ajuste a uma linguagem particular ou domínio analisado.

Na primeira etapa, o *pipeline* de pré-processamento converte os documentos de entradas em um conjunto de regras ou camadas lingüísticas representados em arquivos XML separados. Em seguida, a detecção de menções verifica essas regras ou camadas para extrair e atribuir propriedades aos termos (número, gênero, etc.). O próximo passo, extração de características, descreve os pares de menções $\{Mi, Mj\}, i < j$ como um conjunto de características. Com os pares de conjunto de características definido, o decodificador gera uma base de exemplo por meio de um processo de seleção da amostra e aprende a classificar tais pares. Por fim, o

[7]http://http://bart-coref.org/

codificador gera testes de exemplos distintos da fase anterior, executa o classificador e particiona esses conjuntos de pares como coreferências. BART utiliza-se das técnicas citadas como part-of-speech, corpos lingüísticos e técnicas de aprendizagem de máquina [17].

Após a execução de todas essas etapas, o resultado final é um arquivo XML como da Figura 2.7. O XML é bastante intuitivo, pois segue as abreviações utilizados no processamento de linguagem natural: a tag $< s >$ significa uma nova sentença ou paragráfo; a tag $< wpos =$"atributo"$>$ mostra a classificação lingüística do termo (substantivo, preposição, adjetivo, etc); e quando ocorre uma coreferência a tag $< coref >$ é utilizada, essa tag sempre é seguido pelo atributo *set-id*, o qual é um número único para aquela entidade. Observe que na Figura 2.7, a entidade Nokia possui como identificador o valor *set_22*, com isso sempre que houver referência a ela, esse valor será usado na propriedade *setid*.


Figura 2.7: Exemplo de um arquivo XML como saída da execução da ferramenta BART.

## 2.3 Avaliação

Para avaliar a qualidade dos classificadores utilizados neste trabalho, utilizamos a métrica de acurácia[19, 2, 11]. Ela consiste na fração de documentos de teste que foram assinalados corretamente pelo classificador. Portanto, a acurácia verifica se os documentos, para os quais já se conhece as classes a qual pertencem, foram classificados corretamente. A fórmula para o cálculo é $Acc(cp) = (TA)/(TC)$, onde TA é o total de documentos corretamente classificados e TC o número total de documentos na coleção.

A acurácia final de cada classificador foi determinada como a média das acurácias observadas em cinco divisões dos dados em treino e teste. Estas divisões foram obtidas através de um procedimento estatístico conhecido como validação cruzada de cinco partições[19]. Essa técnica consiste em dividir a coleção em cinco partições. Em seguida são realizadas cinco rodadas de experimentos de forma que na i-ésima rodada, a i-ésima partição é usada como conjunto de teste enquanto as demais são usadas como conjunto de treino. Desta forma, a acurácia obtida em cada rodada nunca é sobre um documento previamente avaliado. A acurácia final é tomada como a média das cinco rodadas.

Finalmente, em todas as comparações entre dois métodos, para termos certeza que a diferença observada em seus desempenhos não é produto de uma escolha casual de dados, realizamos o teste-T[19] para estabelecer a confiança estatística da diferença observada. Tal métrica é utilizada para verificar se há uma hipótese nula quando a estatística de teste segue uma distribuição-$t$ de Student. Em particular, nossa hipótese nula é que os métodos comparados tem desempenho similar (empate). Em particular, o teste-T mede a probabilidade $p$ de que a diferença observada supera a hipótese nula.

Em todas as comparações entre métodos relatadas nesse trabalho, iremos identificar com um asterisco (‘*’) quando o valor observado é significativo considerando um nível de 95% (ou seja, a chance da hipótese nula ser correta é inferior a 5%). Em

todas as discussões de resultados, vamos sempre levar em conta a validade estatística observada.

## 2.4 Trabalhos Relacionados

A análise de sentimentos tem sido empregada em diversos domínios, como resenhas de filmes e produtos. Embora métodos de análise possam ser baseados puramente em técnicas léxicas (a polaridade do texto é inferida da polaridade intrínseca da palavra), métodos baseados em conjuntos mais ricos de atributos que capturam léxico, contexto e outros elementos do texto são mais comumente usados.

Um dos primeiros trabalhos na linha de reconhecimento de polaridade foi proposto em [14], que demonstrou que a classificação de documentos de acordo com a sua polaridade é similar à classicação com base em seus tópicos. Os experimentos foram realizados com base nas resenhas de filmes do site *Internet Movie Database* (IMdB), em que apenas os documentos que continham uma nota associada ao texto do documento foram utilizados. Nesse trabalho, os autores usaram os classificadores Naive Bayes, Máxima Entrópia e Support Vector Machine (SVM) [19]. Destes, o Naive Bayes apresentou o pior desempenho. Os demais classificadores tiveram desempenho comparável ao realizado por pessoas. Os resultados apresentados demonstram que a classificação dos documentos por sua polaridade é tão desafiante quanto por tópicos, uma vez que é comum o uso de ironia e o contexto é importante para a definição da polaridade, já que palavras de cunho positivo podem ocorrer em frases negativas (o contexto) ou vice-versa.

Para lidar com o problema de contextualização propuseram em [18] uma abordagem que explora características de frases para analisar o sentimento dos textos. A base de dados usada foi *Multi-perspective Question Answering* (MPQA) cujo conteúdo consiste de documentos da língua inglesa, com conteúdo detalhado e forte significado emocional. O principal resultado demonstrado é que um conjunto léxico

pré-classificado como positivo e negativo a priori não funciona sempre, pois depende do contexto em que o termo é utilizado. Por exemplo, no segmento de texto na Tabela 2.1, o termo Trust expressa um sentimento positivo. Porém, no contexto dado, a palavra não é usada para expressar um sentimento e sim, o título da entidade. Os autores desse trabalho também observaram que termos classificados como positivos e negativos são comuns em frases neutras.

| |
|---|
| *Philip Clapp, president of the National Environment Trust, sums up well the general thrust of the reaction of environmental movements: "There is no reason at all to believe that the polluters are suddenly going to become reasonable* |

Tabela 2.1: Exemplo de termo positivo usado em contexto negativo.

Diferente dos dois trabalhos citados, [20] demonstrou que a polaridade de um documento deve ser obtida com base em diferentes tópicos dentro do mesmo texto. Este trabalho é particularmente interessante para nós, uma vez que também consideramos que a polaridade deva ser tomada a partir de segmentos de texto.

O autor em [1] foi um dos primeiros que propôs o uso de análise de polaridade no domínio financeiro, motivado pela possibilidade de previsão das reações do mercado de ações. O autor usou como base de dados a *Reuters Key Developments Corpus*, que contém notícias no período de 1998 a 2009. Das companhias nesta coleção, o autor utilizou somente as com mais de 20 notícias. Seus resultados foram obtidos com o uso de técnicas de processamento de linguagens naturais e classificadores baseados em Árvores de Decisão e SVM. O autor concluí que é possível aprender os termos mais usados e de maior impacto para medir polaridade, com desempenho tão bom quanto o de avaliadores humanos. Ele também observou que os modelos aprendidos no domínio financeiro não foram úteis quando aplicados a outros domínios.

Em [4], os autores estudaram a relação entre a polaridade de textos associados a companhias e o seu desempenho no mercado de ações. O *Twitter* foi usado como base para experimentação e *Google-Profile of Mood States* (GPOMS) para valores de polaridade, o autor tem como objetivo antecipar o mercado de ações financeiras. Ao estudar um grande volume de dados do *Twitter*, os autores observaram que mudanças

no estado emocional do público tem impacto dias depois no mercado financeiro. Seus resultados foram alcançados com simples técnicas de processamento de linguagem natural.

Outros trabalhos que exploraram a relação entre polaridade e desempenho no mercado de ações foram propostos por [8] e [9]. Estes trabalhos se basearam na categorização das emoções básicas do homem, segundo Darwin: raiva, medo e tristeza, entre outras. Também delimitaram os sentimentos de acordo com múltiplas dimensões ao invés de categorias discretas. Duas dimensões primárias foram utilizadas: um eixo bom-mal e outro eixo de forte-fraco. Os experimentos foram realizados com bases em notícias e comportamento do mercado de ações relativos a duas companhias aéreas da Irlanda. A técnica para mensurar a polaridade do texto consistiu em construir um grafo que representa o texto todo. Neste grafo, os nós são termos do texto com seus respectivos valores de polaridade (obtidos com a ferramenta SentiWordNet, que mede os termos em positivos e negativos de acordo com a WordNet). Em suma, sua abordagem baseia-se no uso de um conjunto léxico com termos positivos e negativos, com o apoio da teoria de Darwin que tenta identificar quão intenso é esse sentimento. Para os termos positivos houve uma alta revocação, porém uma baixa precisão. Aos termos negativos ocorreu uma alta precisão, entretanto uma baixa revocação. Por fim, o autor concluí que o mapeamento direto dos termos do texto com os termos da teoria de Darwin não é algo simples e claro de ser feito.

Finalmente, [16] também apresenta um estudo sobre o impacto da polaridade na previsão financeira. Os autores analisam notícias financeiras com base em diferentes representações textuais: *bag of words*, sintagmas nominais e nome de entidades. Neste trabalho, os autores observaram que há uma correlação entre o preço futuro de uma ação e os seus preços tomados no momento em que artigos sobre ela são publicados, quando considerados em conjunto com as polaridades dos seus termos presentes nestes artigos.

Uma característica comum dos trabalhos citados anteriormente é que eles consideram que a polaridade é atribuída ao documento e não às entidades citadas nele. Enquanto este é o caso pra textos que representam a opinião sobre um certo tema, produto ou serviço, não é o caso para documentos desestruturados que citam várias entidades e nos quais estamos interessados na polaridade de todas elas. O único trabalho na literatura que encontramos, focado na análise de polaridade para múltiplas entidades em textos mais longos, foi o proposto em [12]. Neste trabalho os autores apresentam um arcabouço para a modelagem de polaridade dentro de sub-contextos. Como a polaridade de cada contexto atômico está associado a uma certa entidade, ao compor um escore de polaridade para um contexto maior, naturalmente polaridades são obtidas por entidade. Esta estratégia é chamada de composicional. Os autores consideram esta tarefa de análise multi-entidade em múltiplos contextos como uma tarefa nova para o qual obtiveram acurácia apenas um pouco inferior ao de anotadores humanos. Além disso, o domínio não é restritivo, ou seja, a abordagem é aplicável a diversos domínios. Embora o código fonte relacionado com o método proposto não esteja disponível, uma ferramenta online possibilita a análise de polaridade para um número limitado de documentos por dia. Contudo, os autores não responderam a nenhuma das nossas solicitações de acesso à ferramenta. Adicionalmente, não foi possível implementar o método proposto pelos autores uma vez que ele não é descrito em um nível de detalhe que permita a sua implementação e depende de uma gramática de sentimentos que não é disponibilizada publicamente.

O nosso trabalho é diferente de todos os citados ao focar na questão de análise de polaridade para documentos que citam múltiplas entidades, nos quais estas entidades são o foco da polaridade. Esses documentos constituem textos de notícias no domínio financeiro. Nesta pesquisa, usamos o método de [1] como base para o cálculo de polaridade, uma vez que ele é bastante efetivo no cenário de polaridade única.

# Capítulo 3

# Aprendizado de Polaridade para Múltiplas Entidades

Neste capítulo apresentamos as nossas propostas para a segmentação de texto de acordo com a entidade referenciada.

## 3.1 Aprendizado de Múltiplas Polaridades como um Problema de Múltiplos Modelos

Seja $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ uma coleção de documentos e $y_i \in \{+, -, N\}$ a polaridade do documento $d_i$. A tarefa de análise de polaridade pode ser vista como uma tarefa de classificação onde, para cada documento $d_i$, pretende-se predizer o rótulo $y_i$, ou seja, encontrar uma função (modelo) $f : \mathcal{D} \implies \{+, -, N\}$, tal que $f(d_i) = y_i$.

No problema de análise de polaridade com $m$ entidades $\{E_1, E_2, ..., E_m\}$, para cada documento $d_i$ temos um conjunto de polaridades $Y_i = \{y_{i1}, y_{i2}, ..., y_{im}\}$. Assim, a polaridade $y_{ij}$ corresponde à polaridade do documento $d_i$ com relação à entidade $E_j$. Neste caso, a tarefa de análise de polaridade pode ser vista como uma classificação em que, dada a coleção de documentos $\mathcal{D}$, pretende-se encontrar $m$ funções $f_j : \mathcal{D} \implies \{+, -, N\}$, tal que $f_j(d_i) = y_{ij}$.

Note que documentos em que a entidade $E_j$ não ocorre, não há contribuição para o aprendizado da função $f_j$ uma vez que estes documentos não podem ser usados como exemplos de casos positivos, negativos ou neutros pra $E_j$. Assim, de fato, a função $f_j$ é melhor representada como $f_j : \mathcal{D}_j \implies \{+, -, N\}$, onde $D_j$ é o subconjunto de documentos que referenciam a entidade $E_j$.

Da mesma forma, fragmentos de um documento que se referem a uma entidade $E_j$ provavelmente não deveriam ser usados como exemplos de treino para uma entidade $E_k, k \neq j$. Imagine que o fragmento (ou o documento como um todo) fosse avaliado como negativo. Não parece adequado usar esse fragmento como exemplo de negativo para $E_k$ se ele não se refere a $E_k$. Por exemplo, no documento exibido na Figura 3.1, o primeiro parágrafo faz referência apenas à Nokia. Logo, é questionável que ele devesse ser usado como exemplo de treino para Apple, Google e Microsoft, mesmo que estas três companhias sejam citadas em outros parágrafos no texto.

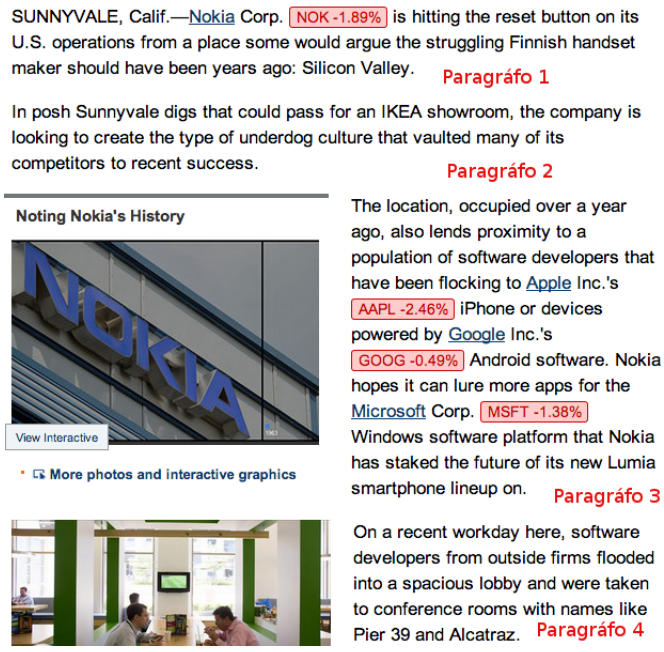


Figura 3.1: Texto de um documento com entidades Nokia, Apple, Microsoft e Google.

Desta forma, seja $d_i^j$ o documento formado por todos as sentenças de $d_i$ que fazem referência à entidade $E_j$. Nós denotamos por $\mathcal{D}^{(j)}$ o conjunto de todos os

documentos $d_i^j$, ou seja, o conjunto de todos os documentos que são formados apenas por sentenças que fazem referência a $E_j$. Dadas estas definições, nós podemos re-escrever a função $f_j$ como $f_j : \mathcal{D}^{(j)} \implies \{+, -, N\}$.

Com base no discutido nesta seção, podemos definir duas estratégias para o aprendizado de múltiplas polaridades usandos múltiplos modelos:

- Modelo baseado em Documentos (MD): obtenha cada função $f_j$ a partir da coleção de documentos $\mathcal{D}_j$, ou seja, o conjunto de *documentos* que referenciam $E_j$.

- Modelo baseado em Sentenças (MS): obtenha cada função $f_j$ a partir da coleção de documentos $\mathcal{D}^{(j)}$, ou seja, o conjunto de documentos formados por *sentenças* que referenciam $E_j$.

Note que para a estratégia MD, determinar o conjunto de documentos associados a uma entidade $E_j$ é trivial, uma vez que consiste em observar em que documentos a entidade $E_j$ ocorre. Este não é o caso para a estretégia MS, uma vez que não é tão simples determinar que sentenças se referem à entidade $E_j$. Na próxima seção, discutimos formas de determinar que sentenças citam uma entidade $E_j$.

## 3.2 Mapeamento de Sentenças e Entidades

Como descrito anteriormente, a Figura 3.1 apresenta um documento em que são citadas as entidades Nokia, Apple, Google e Microsoft. É relativamente simples observar que o primeiro parágrafo se refere a Nokia e o terceiro se refere às quatro entidades, uma vez que elas são citadas diretamete nos mesmos. Dos dois parágrafos restantes, o segundo também se refere a três das quatro entidades enquando o último não cita nenhuma delas.

No segundo parágrafo, as entidades são citadas indiretamente. O termo “the company” é usado para se referir a Nokia, enquanto “competitors” corresponde a

Apple e Google. Este tipo de citação indireta (anáfora) não é tão simples de capturar pois exige uma compreensão mais profunda do texto (compreender que "company" se refere à companhia que é o foco do discurso na sentença em que "company" é usado) e, algumas vezes, conhecimento de domínio e contexto (saber que na época em que o texto foi escrito, a Nokia competia com Apple e Google, mas não com Microsoft).

Assim, a tarefa de determinar a que entidades se referem as sentenças pode ser realizada usando desde heurísticas simples de ocorrência no texto (a primeira sentença se refere a Nokia pois a string "nokia" ocorre no texto) a complexas estratégias que envolvem métodos sofisticados de processamento de linguagem natural, como a resolução de anáforas (o segundo parágrafo se refere à Nokia pois "company" corresponde à Nokia).

Com base nessas observações, propomos seis variantes para estratégia MS, três delas baseadas em simples heurísticas de ocorrência de termos e três delas baseadas em resolução de anáfora. As seis variações são descritas a seguir:

- Estratégia MS1: Parágrafo/sentença é atribuído à entidade cujo nome ocorre nele. Se nenhuma entidade está presente, ele é atribuído a todas. Esta heurística tenta capturar situações como a do segundo parágrafo da Figura 3.1;

- Estratégia MS2: Parágrafo/sentença é atribuído à entidade cujo nome ocorre nele. Se nenhuma entidade está presente, ele é descartado. Esta heurística tenta capturar situações como a do quarto parágrafo da Figura 3.1;

- Estratégia MS3: Parágrafo/sentença é atribuídos à última entidade cujo nome foi citado, se o nome de nenhuma entidade ocorre nele. Esta heurística se baseia na ideia de que se nenhuma nova citação foi feita, o texto provavelmente continua se referindo à última entidade citada;

- Estratégia MS4: Parágrafo/sentença é atribuído à entidade referenciada nele. Se nenhuma entidade está referenciada, ele é atribuído a todas. Equivalente

de MS1 com resolução de anáfora;

- Estratégia MS5: Parágrafo/sentença é atribuído à entidade referenciada nele. Se nenhuma entidade está referenciada, ele é descartado. Equivalente de MS2 com resolução de anáfora;

- Estratégia MS6: Parágrafo/sentença é atribuído à última entidade referenciada. Equivalente de MS3 com resolução de anáfora;

Note que nas estratégias MS1 a MS3, a noção de citação corresponde à verificação da ocorrência do nome da entidade no texto, como é o caso de Nokia no primeiro parágrafo da Figura 3.1. Para este caso, o texto é processado usando as técnicas de processamento linguagem natural descritas a seguir. Em primeiro lugar, o texto é segmentado em sentenças/parágrafos, que são definidos como sequências de tokens (conjuntos de caracteres que correspondem a palavras) terminados com ponto ou ponto seguido de quebra de linha. Em seguida, todas as *stopwords* (preposições, artigos, numerais) são removidas do texto. Após isso, os nomes das entidades são casados com o texto.

Para exemplificar o resultado das estratégias para cada entidade, o texto da Figura 3.1 contém 4 paragráfos apenas. Para a estratégia MS1, o único paragráfo que não se repetirá para Google, Apple e Microsoft é o primeiro paragráfo, isso porque de acordo com a definição, a única entidade citada nele é a Nokia. Essa estratégia é a que mais se assemelha com o texto todo do documento original (usando na estratégia MD). Portanto, para a Nokia, os paragráfos 1, 2, 3 e 4 serão os fragmentos de texto referentes a ela. Para as entidades Apple, Google e Microsoft o paragráfo 2, 3 e 4 são os seus fragmentos de texto. Nota-se que o paragráfo 1 é descartado para essas entidades, pois ele já cita outra entidade (Nokia).

Com a Estratégia MS2, a diferença consiste em descartar o parágrafo em que nenhuma entidade está presente. Ou seja, para a Nokia apenas os paragráfo 1 e 3 serão os fragmento de texto. As demais entidades terão como texto do documento

o paragráfo 3.

Por fim, na Estratégia MS3 o paragráfo é da última entidade citada. Com isso, o parágrafo 2 é atribuído à Nokia e o parágrafo 4 à Microsoft. A Nokia ainda é citada nos parágrafos 1 e 3, enquanto as outras, no parágrafo 3.

Para as estratégias MS4 a MS6, para considerar que um páragrafo cita uma entidade é necessário que haja uma referência à entidade no parágrafo, mesma que indireta. É o caso de Nokia no segundo parágrafo da mesma figura. Para este caso, a ferramenta BART é usada para resolver as anáforas encontradas no texto.

Como exemplificado no Capítulo 2.4, o BART tem como saída um documento XML em que para cada entidade referenciada é definido um conjunto de identificadores únicos. A Figura 3.2 apresenta um trecho da representação do documento de acordo com o BART:


Figura 3.2: Representação do BART para o documento.

Para a implementação das estratégias, o texto do documento torna-se o conteúdo das tags do arquivo XML gerado. Na Figura 3.2, a Nokia possui o id igual a set_6 e set_13, enquanto a Apple o id é set_1. Além disso, a verificação do fim de um paragráfo passa a ser o elemento <\s >, como exemplificado no Capítulo 2.4.

Para a Estratégia MS4, mesmo que o termo Nokia apareça no primeiro paragráfo, o BART limitou sua referência ao seguinte trecho: *SUNNYVALE, CALIF. - Nokia.*. Em seguida, o próximo trecho relativo à Nokia é no terceiro paragráfo: *Nokia hopes it can lure more apps for the Microsoft Corp.; Windows software plataform that Nokia has staked the future of its new Lumia smartphone lineup on.* Note que não é necessário aparecer o nome da entidade, basta que haja a tag *<coref set-id="set_6"><\coref>* ou *<coref set-id="set_13"><\coref>* com os id's atribuídos a elas. Portanto, esses dois trechos fazem parte do novo conteúdo do documento, além dos paragráfos em que não há nenhuma referência a nenhuma entidade alvo. O que faz com que o novo documento seja parecido com o original, entretanto diferente das abordagens MS1 a MS3, não é todo o paragráfo que é inserido, mas trechos em que BART entende como novo paragráfo, por meio da tag <s><\s>.

A diferença entre as Estratégias MS4 e MS5 para o BART corresponde à verificação de alguma entidade referenciada entre as tags <s><\s>. Caso não haja, o trecho é descartado. Por fim, a Estratégia MS6, verifica a última entidade referenciada entre as tags. Nesse caso, é possível que o BART entenda que mesmo que uma entidade apareça na tag, ela não seja referenciada.


Figura 3.3: Exemplo de excessão para Estratégia 3.

A Figura 3.3 mostra que a Nokia está sendo referenciada no paragráfo como *<coref set-id="set_6"><w pos="nn">nokia<\w><\coref>*, mas a Microsoft que também é citada no paragráfo, não foi marcada como referênciada pelo BART.

Nesses casos, perderiamos informações para Microsoft, o que poderia resultar em um documento vazio para ela. Para evitar isso, nesses casos o trecho é atribuido à entidade cujo nome ocorre no trecho, como ocorre nos métodos MS1 a MS3.

# Capítulo 4

# Experimentos

Neste capítulo detalharemos os experimentos realizados para avaliar as estratégias propostas no Capítulo 3.

## 4.1 Metodologia Experimental

A primeira etapa da avaliação consistiu na criação de uma base de documentos financeiros em língua inglesa. Como nosso objetivo inicial era criar uma coleção que, mais tarde, fosse usada não apenas para análise de polaridade, mas também para previsões, foram coletadas cerca de 60 mil documentos ao longo de quatro meses. Esses documentos foram coletados dos seguintes sites: *Bloomberg*[1], *The New York Times*[2], *Financial Times*[3], *Reuters*[4], *Forbes*[5], *AllThingsD*[6] e *Cnn Money*[7].

Com o conjunto de entidades alvo pré-determinado, os coletores foram implementados para recuperar apenas páginas que citavam tais entidades. Para isso, o coletor preenchia o formulário de busca interna do site, citados anteriormente, com o nome de cada entidade alvo e extraía as *urls* de notícias que o sistema de busca

[1] `http://www.bloomberg.com`
[2] `http://www.nytimes.com`
[3] `http://www.ft.com`
[4] `http://www.reuters.com`
[5] `http://www.forbes.com`
[6] `http://allthingsd.com`
[7] `http://money.cnn.com/`

retornava. A Figura 4.1 mostra: a página inicial pela busca da entidade Apple no sítio da Reuters e a página de resultados dessa mesma busca.

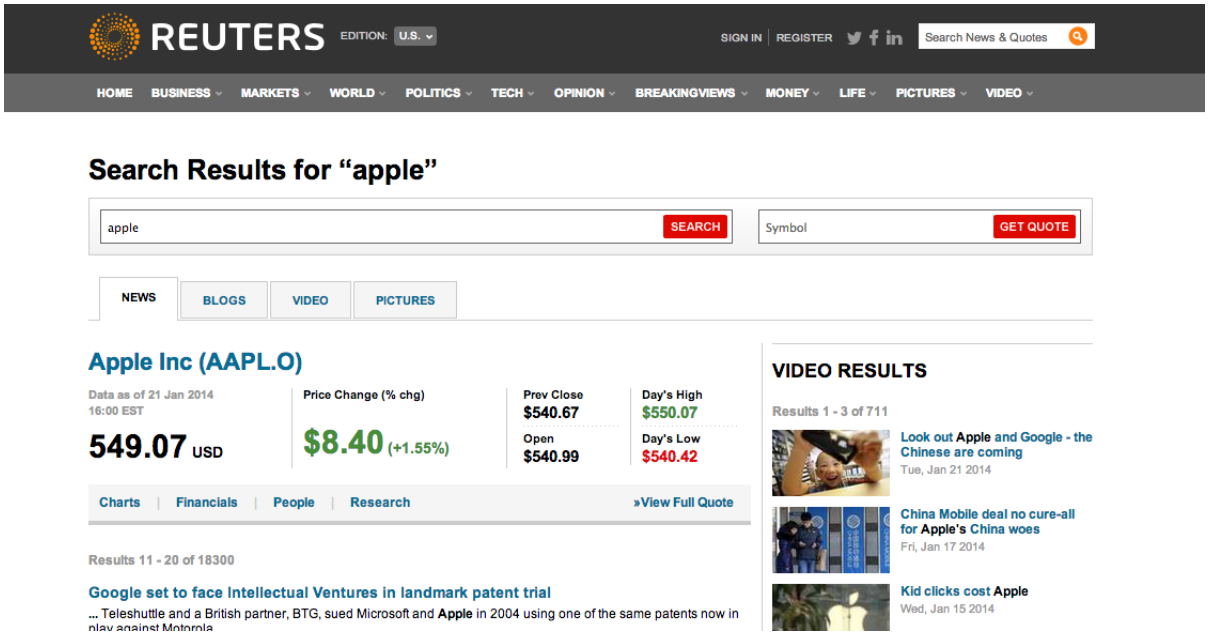


Figura 4.1: Preenchimento da busca interna da Reuters por um coletor.

Além disso, notou-se que a busca interna por entidades distintas poderia retornar *urls* iguais mesmo que o termo da busca fosse distinto. Para evitar a coleta da mesma página por várias vezes, a solução proposta foi a verificação da url ao longo da execução do coletor pela página da entidade em questão e das *urls* já coletadas pelas demais entidades. Isso ocorre quando há mais de uma entidade pertencente ao conjunto no mesmo documento. Tabela 4.1 mostra o total de páginas coletadas referente ao site visitado.

| Site | Páginas Coletadas |
|---|---|
| Reuters | 3.006 |
| AllThingsD | 15.375 |
| New York Times | 1.168 |
| CNN Money | 3.434 |
| Forbes | 34.653 |
| Financial Times | 113 |
| Bloomberg | 3.968 |
| Total | 61.717 |

Tabela 4.1: Sites coletados com o total de páginas.

Após a criação da coleção de documentos, o próximo passo foi classificar as páginas por usuários entre positivo, negativo e neutro. Contudo, como a coleção contém mais de 60 mil páginas, era inviável a rotulação manual de todas elas. Então, um subconjunto aleatório de 1.000 páginas foi selecionado para ser rotulado por cerca de 40 usuários. O pré-requisito para o usuário participar foi ser capaz de ler em língua inglesa, uma vez que todas as páginas coletadas estavam em inglês. Cada usuário recebeu um lote de 25 páginas para realizar a rotulação.

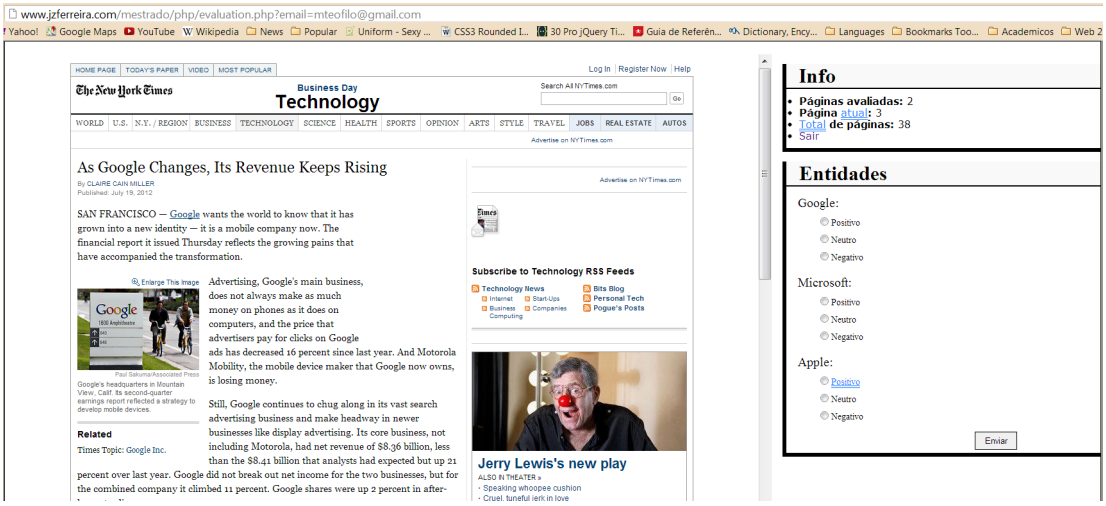


Figura 4.2: Sistema de avaliação.

Para isso, um sistema de avaliação foi desenvolvido, no qual a página com a notícia é apresentada ao usuário e juntamente com a opção de classificar as entidades citadas no texto, dentro do conjunto das cinco companhias selecionadas. A Figura 4.2, mostra a interface do sistema.

Ao fim da avaliação, dos mil documentos, 300 foram avaliados como positivos, 85 como negativos e 615 como neutros. Com relação a cada entidade, a Tabela 4.2 apresenta a distribuição da avalição dos usuários e o total de páginas referente a cada entidade. Note nesta tabela que a distribuíção é muito desbalanceada tanto em termos de polaridade (61% são neutras e apenas 13% são negativas) quanto entidades (a Apple foi citada em cerca de 95% dos documentos enquanto a Nokia em 12%).

A Tabela 4.3 apresenta a distribuição das entidades por páginas, nota-se que o conjunto de páginas constitui páginas que citam pelo menos uma das entidades do conjunto pré-definido. Como podemos notar, a maioria dos documentos (62%) cita mais de uma entidade. Contudo, o número mais comum de entidades por documento é uma (38% dos documentos).

A Tabela 4.4 apresenta a distribuição de polaridades por página. Apesar da maioria dos documentos citar mais de uma entidade por página, em apenas cerca de 30% deles, polaridades distintas foram atribuídas em uma mesma página. Isso implica que em 30% dos documentos, o classificador seria treinado de forma equivocada para, pelo menos, uma entidade. De qualquer forma, para 70% dos documentos, não há vantagem em usar múltiplos modelos, para a amostrada analisada.

Com a coleção de 1.000 páginas classificadas por seres humanos em três classes (positivo, negativo e neutro) e o conjunto de páginas para cada entidade, um extrator extraiu o conteúdo textual da notícia, descartando tags HTML. Os arquivos resultantes foram fragmentados conforme descrito no Capítulo 3, seção 3.2, gerando as coleções correspondentes às estratégias MS1, MS2 e MS3. Também foi executado o BART em todos os 1.000 documentos de forma a gerar as coleções para as estratégias MS4, MS5 e MS6. Além disso, uma variação para a estratégia MD, em que usa-se todo o texto do documento com o rótulo de polaridade da entidade, foi proposto. O MD-G também utiliza-se de todo o texto do documento, porém o rótulo de polaridade da entidade é modificado para o rótulo do documento. Portanto, dado um determinado documento em que a entidade Apple é rotulada como positiva, ela passa a ser neutro visto que a polaridade do documento como um todo é neutro.

| Entidade | POS | NEG | NEUTRO | TOTAL |
|---|---|---|---|---|
| Apple | 261 | 131 | 562 | 954 |
| Google | 105 | 39 | 276 | 420 |
| Samsung | 81 | 58 | 197 | 337 |
| Microsoft | 55 | 31 | 195 | 281 |
| Nokia | 29 | 25 | 71 | 125 |

Tabela 4.2: Distribuição de Polaridade por Entidade.

Com as coleções criadas, o método linha de base foi executado conforme descrito em [1]. Assim, o texto foi normalizado com as seguintes técnicas: (1) aplicação de *stemming* nos *tokens*; (2) remoção das palavras *stopwords* de acordo com a ferramenta *WordNet*; (3) a remoção de palavras que ocorressem menos de três vezes na coleção, isso porque entende-se que o termo tem baixa importância. Em seguida, foi usado o classificador SVM (ferramenta *LIBSVM*) com protocolo de experimentação de validação cruzada de 5 partições (geradas com o WEKA). Este protocolo consiste em separar cada coleção em cinco partes e, em cada rodada entre cinco rodadas, usar uma parte para teste e as demais partes para treino. Os resultados finais são dados como a acurácia média das cinco partições.

O teste de validade estatística de Student (Teste T) foi aplicado em todas as comparações entre os métodos, de forma que reportamos diferenças significativas considerando um nível de significância de 95%.

## 4.2 Resultados

Nesta seção apresentamos os resultados obtidos nos experimentos. Primeiro, mostramos o ganho obtido com o uso de múltiplos modelos para, então, discutirmos o impacto da segmentação de documentos.

### 4.2.1 Múltiplos Modelos

Inicialmente, nós comparamos a estratégia Global com a estratégia MD. Na estratégia Global, o método proposto por [1] foi treinado e testados com todos os

| Entidades | Páginas |
|---|---|
| 1 | 381 |
| 2 | 310 |
| 3 | 164 |
| 4 | 87 |
| 5 | 58 |

Tabela 4.3: Distribuição de Entidades por Páginas.

1000 documentos da coleção (independente das entidades) para aprender as polaridades atribuídas aos documentos. Na estratégia MD, o mesmo método aprendeu cinco modelos, um para cada entidade, baseados nas polaridades atribuídas às entidades.

A Tabela 4.5 mostra os resultados obtidos, considerando os mil documentos e os 617 documentos com entidades distintas. Nesta tabela, as cinco primeiras linhas correspondem à avaliação dos métodos usando as polaridades das entidades.

Como esperado, usar múltiplos modelos, um por entidade, é melhor que usar um modelo global. No caso da coleção de mil documentos, apenas para o caso da Apple, o método MD não apresentou um ganho significativo. Os ganhos foram melhores, em geral, para as entidades com menos documentos. Isso sugere que o aprendizado por entidade foi melhor pra aprender padrões específicos das entidades menos populares, ao escapar dos vícios característicos das entidades mais populares. Por exemplo, a entidade Apple é a mais citada da coleção de 1000 documentos, entretanto diversas citações são referentes apenas por ser a fabricante do Iphone e/ou Ipad, tendo sua polaridade como neutra nesses casos. Os ganhos são bem maiores na coleção de 617 documentos, uma vez que nestes documentos, há múltiplas entidades distintas, o que não é capturado pelo método global. Neste caso, o ganho maior para entidades com menos documentos ainda é mais evidente.

Quando treinado em todos os documentos e avaliado com rótulos dos documentos (uma única polaridade por documento, independente das entidades nele), o método Global apresentou desempenho similar nas duas coleções. Isto indica que a coleção com múltiplas entidades diferentes (617 documentos) não apresenta uma caracterização muito distinta da coleção completa, quando não consideramos entidades

| Polaridades | Páginas |
|---|---|
| 1 | 697 |
| 2 | 257 |
| 3 | 46 |

Tabela 4.4: Distribuição de Polaridades por Páginas.

| | Todos Documentos | | | 617 Documentos | | |
|---|---|---|---|---|---|---|
| Entidades | Global | MD | Ganho | Global | MD | Ganho |
| Apple | 55.87 | 58.49 | 4.68 | 43.22 | 63.68 | 28.55* |
| Google | 54.17 | 61.42 | 13.38* | 45.44 | 61.80 | 36* |
| Samsung | 52.46 | 61.11 | 16.48* | 45.31 | 59.70 | 31.75* |
| Microsoft | 54.42 | 71.54 | 31.45* | 50.36 | 77.43 | 53.75* |
| Nokia | 45.72 | 57.60 | 25.98* | 29.23 | 62.40 | 113.47* |

Tabela 4.5: Método de aprendizado global comparado com estratégia MD considerando todos os documentos e apenas os 617 documentos que apresentam mais de uma entidade do grupo pré-definido. Ganhos estatísticamente significativos estão marcados com um asterisco ('*').

individuais. Assim, o resultado de caracterizações tão distintas entre a avaliação por entidade e por documento para o método Global se torna mais notável na grande perda sofrida pelo método quando avaliado por entidades usando a coleção de 617 documentos (coleção com múltiplas entidades).

Para compreender melhor o impacto do treinamento com rótulos de documentos ou específicos por entidade, a Tabela 4.6 mostra os resultados obtidos para o método Global, o método MD e a sua variante MD-G, considerando a coleção com mil documentos. MD-G é uma variante de MD que foi treinada usando rótulos dos documentos em lugar de rótulos das entidades. Como MD, contudo, MD-G foi treinado usando documentos específicos de cada entidade.

| Entidades | Global | MD-G | MD | Ganho |
|---|---|---|---|---|
| Apple | 55.87 | 58.69 | 58.49 | -0.34 |
| Google | 54.17 | 58.09 | 61.42 | 5.73* |
| Samsung | 52.46 | 63.79 | 61.11 | -4.20* |
| Microsoft | 54.42 | 64.43 | 71.54 | 11.03* |
| Nokia | 45.72 | 71.20 | 57.60 | -19.10* |

Tabela 4.6: Estratégia MD-G foi treinada usando documentos específicos por entidade e rótulos dos documentos. Ganhos foram calculados entre MD e MD-G. Valores estatísticamente significativos estão marcados com um asterisco ('*').

Como podemos observar na Tabela 4.6, a estratégia MD obteve ganhos significativos sobre MD-G para Google e Microsoft, com perda não significativa no caso da Apple. Os ganhos contudo foram bem menores que em relação ao método Global, o que sugere que grande parte do ganho esta associada ao treinamento usando doc-

umentos específicos por entidade. O ganho relacionado com o uso de rótulos de entidade é, portanto, menor. O resultado surpreendente nessa tabela é aquele relacionado com a Nokia. O método MD foi significativamente inferior ao MD-G nesse caso. Uma observação mais detalhada desse resultado mostra que o MD-G foi capaz de aprender melhor neutros pra Nokia (que são frequentes pra outras entidades em documentos em que a Nokia aparece) que o modelo treinado especificamente pra ela. Como neutros são as polaridades mais comuns, isso levou o MD-G a um desempenho inesperadamente melhor.

### 4.2.2 Segmentação de Documentos

Nesta seção, nós avaliamos se vale a pena segmentar os documentos por sentenças atribuindo a cada entidade apenas as sentenças que a referenciam. Para tanto, comparamos todas as seis estratégias de segmentação de sentenças propostas no Capítulo 3 com a estratégia MD, em que documentos não são segmentados.

Foram realizados experimentos tanto com a coleção completa quanto com a coleção com 617 documentos. Os resultados destes experimentos são apresentados nas Tabelas 4.7 e 4.8. Nestas tabelas, as linhas representam as estratégias MD e as estratégias de segmentação em sentenças, baseadas em ocorrência (MS1, MS2 e MS3) e resolução de anáforas (MS4, MS5 e MS6). Os resultados representam a acurácia do classificador proposto por [1] ao atribuir polaridades positiva, negativa e neutra para as entidades Apple, Google, Samsung, Microsoft e Nokia. Para cada método baseado em sentenças, é apresentado o ganho (ou perda) em relação ao método MD.

Observamos na Tabela 4.7 que a estratégia MS1 (sentenças que não referenciam entidades são atribuídas a todas) apresentou ganhos não significativos sobre MD para Google e Samsung (ou seja, duas entidades com mais documentos na coleção), com perdas nos outros casos (sendo a entidade Apple com mais documentos). Contudo, a variante deste método (MS4), baseada em resolução de anáforas, surpreendente-

| Métodos | Apple | G% | Google | G% | Samsung | G% | MS | G% | Nokia | G% |
|---|---|---|---|---|---|---|---|---|---|---|
| MD | 58.49 | - | 61.42 | - | 61.11 | - | 71.54 | - | 57.60 | - |
| MS1 | 57.96 | -0.89 | 61.66 | 0.38 | 62.58 | 2.39 | 70.11 | -1.98 | 55.20 | -4.16 |
| MS2 | 62.58 | 6.99* | **64.28** | 2.32* | 63.79 | 4.37* | 69.06 | -3.45 | 60 | 11.11* |
| MS3 | 58.49 | 0 | 62.61 | 1.93 | 59.94 | -1.91 | 69.75 | -2.49 | 60 | 4.16 |
| MS4 | 57.96 | -0.90* | 64.28 | 4.65* | 62.28 | 1.91 | 71.54 | 0 | 60 | 4.16 |
| MS5 | **63.52** | 8.60* | **63.09** | 2.71 | **62.88** | 2.89* | 67.63 | -5.46 | 61.60 | 6.94* |
| MS6 | 60.48 | 3.40* | 64.04 | 4.26 | 61.11 | 0.01 | 68.69 | -3.93 | 57.60 | 0 |

Tabela 4.7: Métodos baseados em sentenças (MS1, MS2, MS3, MS4, MS5 e MS6) versus método baseado em documentos (MD), avaliados considerando todos os documentos. Ganhos (coluna G%) estatísticamente significativos estão marcados com um asterisco ('*').

mente foi melhor que MD para a Google, Samsung e Nokia. Nesse caso, o uso de técnicas de resolução de anafóras vale à pena. A pequena diferença dos resultados obtidos por esta técnica em relação ao método MD pode ser atribuída ao fato de que esta é a técnica que menos modifica a coleção de exemplos em relação à coleção original.

Nos resultados para as estratégia MS2 (sentenças que não referenciam entidades são descartadas) e sua variante baseada em anáforas (MS5), observamos um resultado mais satisfatório. Ambas as técnicas conseguiram ganhos sobre as duas entidades mais populares na coleção, com a estratégia MS5 também tendo ganho para a Nokia. Para a Microsoft, houve perda. Ao contrário do caso anterior, aqui a técnica de resolução de anáforas contribuiu para um nível melhor de acurácia do classificador, de fato, entre os melhores obtidos neste trabalho. Ainda assim, dado o custo do processamento de anáforas, o ganho sobre o método MS2 foi discreto.

Nos resultados com as técnicas MS3 e MS6 (sentenças que não fazem referências são atribuidas à última entidade citada, se houver), os resultados foram todos empates e ganho não significativo, exceto pelo caso da Microsoft em que ambas as estratégias houve perda. Ao contrário dos demais casos, este é um pouco mais difícil de interpretar, uma vez que esperávamos um desempenho entre os dois outros pares de estratégias estudados.

De forma geral, deste conjunto de experimentos, podemos concluir que os métodos

de segmentação produzem ganhos relativamente pequenos (da ordem de 5%). O melhor método baseado em anáforas obteve um ganho um pouco maior que o melhor método baseado em ocorrências, o que não justifica o alto custo da técnica empregada para resolver anáforas. Embora alguns resultados positivos foram obtidos para Microsoft e Nokia, a quantidade de documentos destas entidades na coleção não foi grande o suficiente para observamos ganhos significativos.

| Métodos | Apple | G% | Google | G% | Samsung | G% | MS | G% | Nokia | G% |
|---|---|---|---|---|---|---|---|---|---|---|
| MD | 55.56 | - | 61.80 | - | 59.70 | - | 77.43 | - | 62.40 | - |
| MS1 | 66.07 | 18.91* | 65.61 | 6.16 | 69.47 | 16.37* | 75.89 | -1.97 | 60.80 | -2.56 |
| MS2 | 64.22 | 15.58* | 68.95 | 11.56* | 73.31 | 22.80* | 75.12 | -2.98 | 69.60 | 11.53* |
| MS3 | 66.59 | 19.38* | 67.51 | 9.23* | 68.59 | 14.88* | 74.35 | -3.97 | 65.60 | 5.12 |
| MS4 | 65.39 | 17.68* | 69.42 | 12.33* | 69.76 | 16.86* | 78.17 | 0.96 | 66.40 | 6.41 |
| MS5 | 65.73 | 18.29* | 68.70 | 11.17* | 68.27 | 14.35* | 73.20 | -5.45 | 66.40 | 6.41 |
| MS6 | 65.89 | 18.59* | 68.72 | 11.19* | 69.15 | 15.83* | 73.19 | -5.47 | 61.60 | -1.28 |

Tabela 4.8: Métodos baseados em sentenças (MS1, MS2, MS3, MS4, MS5 e MS6) versus método baseado em documentos (MD), avaliados considerando os 617 documentos com entidades diferentes. Ganhos (coluna G%) estatísticamente significativos estão marcados com um asterisco ('*').

Na Tabela 4.8 comparamos os mesmos métodos considerando apenas os 617 documentos que possuem entidades distintas do grupo pré-definido. Como esperado, o número de resultados significativos é bem maior devido ao pequeno número de documentos relativos a todas as entidades. Assim, resultados não significativos foram observados apenas para Microsoft. Os resultados apresentados para a Microsoft foram surpreendentes visto que deveriam apresentar uma melhora pelo menos na estratégia MS2 como todos os demais. Estes resultados reforçam nossa conclusão que o uso de técnicas de segmentação do texto e a sua classificação de polaridade por entidades é melhor do que classificar por documento.

# Capítulo 5

# Conclusões

Neste capítulo, apresentamos as conclusões do nosso trabalho, o que inclui limitações da pesquisa realizada e direções futuras que podem ser exploradas.

## 5.1 Resultados Obtidos

Neste trabalho, investigamos como aprender polaridades para diferentes entidades por meio da construção de múltiplos modelos. Em particular, avaliamos como a segmentação da coleção em documentos relacionados com as entidades e em sentenças relacionadas com a entidade impactam na acurácia do classificador de polaridade. Entre os métodos de segmentação estudados, consideramos simples heurísticas de ocorrência do nome da entidade bem como heurísticas mais sofisticadas baseadas em resolução de anáforas.

Em geral, observamos que a perda ao desconsiderar o aprendizado por entidades é alto em documentos financeiros devido ao número relativamente grande de documentos com múltiplas entidades, mesmo que poucos (cerca de 30% em nossa coleção) desses documentos apresentem diferentes polaridades para as entidades que eles citam. Modelos específicos por entidades (treinados ou não com rótulos específicos por entidade) são capazes de capturar particularidades destas entidades que se perdem quando consideramos coleções com múltiplas entidades e cujas ocorrências são

desbalanceadas.

Também observamos que a segmentação dos documentos em sentenças produz ganhos satisfatórios (da ordem de 5%) em relação a usar os documentos originais. Entre os métodos de segmentação em sentenças, os melhores foram os baseados em resolução de anáforas.

Assim, concluímos que em situações onde é necessário aprender polaridades para diversas entidades, a estratégia simples de aprender um modelo para cada entidade, representada pelos documentos que a citam, com segmentações adicionais produz resultados satisfatórios.

## 5.2 Limitações

A nossa coleção foi formada especificamente por documentos da área financeira. Embora as conclusões que alcançamos possam ser válidas pra outros domínios, se faz necessário a avaliação de outras coleções.

A coleção usada também foi relativamente pequena e muito desbalanceada, dificultando a observação de dados significativos pra todas as entidades estudadas. Nós também focamos em um número pequeno e pré-determinado de entidades quando, em um cenário real, o número de entidades deve ser maior e determinado de forma automática ou ser baseado em uma seleção inicial maior.

Um outro problema observado foi que nossa coleção possuia um grande número de documentos em que nenhuma opinião era emitida em relação às entidades citadas (polaridade neutra pra todas as entidades).

## 5.3 Trabalhos futuros

Na época em que nossa coleção foi coletada, algumas relações eram bem conhecidas entre as entidades que foram alvo da nossa pesquisa. A Microsoft e a Nokia constituiam uma cooperação que, posteriormente, levaria à aquisição da Nokia pela

Microsoft (o que ocorreu após a época em que os dados foram coletados). Nokia e Apple eram rivais do Google, uma vez que eram algumas das poucas fabricantes de smartphones que não adotavam o sistema operacional Android. Por sua vez, a Samsung era a maior parceira do Google entre as fabricantes de celulares. A Apple havia ganho um processo judicial contra a Samsung, relacionada com violação de patentes. É interessante notar que todas estas relações se tornam evidentes em nossa coleção de 303 documentos, quando observamos as proporções de documentos com polaridades em comum entre pares de entidades, excluídos os casos onde uma ou ambas as entidades têm polaridade neutra.

| | Nokia | Google | Microsoft | Apple | Samsung |
|---|---|---|---|---|---|
| Nokia | - | 29 | 78 | 33 | 35 |
| Google | 29 | - | 41 | 39 | 75 |
| Microsoft | 78 | 41 | - | 31 | 42 |
| Apple | 33 | 39 | 31 | - | 33 |
| Samsung | 35 | 75 | 42 | 33 | - |

Tabela 5.1: Proporção de documentos cuja a polaridade é a mesma para o par de entidades considerada (ambas positivas ou ambas negativas).

Estes dados são exibidos na Tabela 5.1. Nela, podemos observar que documentos positivos pra uma entidade tendem a ser negativos pra entidades concorrentes (ver pares Samsung-Apple e Google-Apple). Da mesma forma, entidades que colaboram entre si tendem a ter mais polaridades em comum que diferentes (ver pares Nokia-Microsoft e Google-Samsung). Como relações entre entidades (cooperação e concorrência) podem ser obtidas de bases de conhecimento disponíveis na Web (como a Wikipédia) ou da própria coleção de treino, esta fonte de informação pode ser usada no processo de aprendizado.

Uma forma de utilizar essa evidência no reconhecimento da polaridade seria tratar o problema como uma tarefa de *regressão multi-variada com saídas correlacionadas* em lugar de uma classificação. Neste caso, o alvo da previsão seria uma valor numérico no intervalo de -1 a +1 (onde zero indica uma polaridade neutra) para cada entidade. A técnica empregada deve assumir que as polaridades entre as entidades são dependentes, ou seja, se influenciam mutuamente. No futuro, pre-

tendemos investigar esta e outras ideias no processo de identificação de polaridade para múltiplas entidades. Em particular, pretendemos explorar técnicas como as discutidas em [7], como empilhamento e regressão de posto reduzido.

Também pretendemos investigar como métodos composicionais de detecção de polaridade (baseados na ideia de [12]) podem ser usados para se estimar uma polaridade geral pro documento a partir de polaridades de entidades individuais.

No tocante às limitações do nosso estudo, pretendemos:

- Verificar nossas conclusões em coleções de outros domínios.
- Estudar métodos híbridos baseados em multiplos modelos para lidar com um grande número de entidades. Questões importantes são (a) determinar quando uma entidade deveria ter um modelo específico ou fazer parte de um global e (b) usar modelos baseados em grupos latentes de entidades. A motivação para esta segunda ideia é que entidades de um mesmo setor econômico (ex: Nintendo e Sony) podem ser igualmente afetadas por um mesmo evento/notícia (interesse crescente em jogos em celulares e cada vez menor em consoles).
- Os métodos aqui estudados fazem diferença para documentos em que os autores emitem opinião sobre as entidades, o que não é o caso da maioria dos documentos que observamos na nossa coleção. Assim, pretendemos estudar métodos que detectem opiniões, antes de determinar sua polaridade. Uma primeira aplicação de tais métodos pode ser a coleta de documentos com maior potencial para análise de polaridade.

# Bibliografia


[1] Pablo Azar. *Sentiment Analysis Financial News*. PhD thesis, Hardvard College, 2009.

[2] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009.

[4] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. 1(2):1–8, 2010.

[5] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[6] S.J. Cunningham and P. Denize. A tool for model generation and knowledge acquisition. In *Proc International Workshop on Artificial Intelligence and Statistics*, pages 213–222, Fort Lauderdale, Florida, USA, 1993.

[7] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence in multi-label classification. In *Workshop Proceedings of Learning from Multi-Label Data*, pages 5–12, Haifa, Israel, June 2010.

[8] Ann Devitt and Khurshid Ahmad. A lexicon for polarity: Affective content in financial news text. *Proceedings of Language For Special Purposes*, 2007.

[9] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[10] Pradheep Elango. Coreference Resolution: A Survey.

[11] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[12] Karo Moilanen and Stephen Pulman. Multi-entity sentiment scoring. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *RANLP*, pages 258–263. RANLP 2009 Organising Committee / ACL, 2009.

[13] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.

[14] Bo Pang, Lillian Lee, and 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*, 2002.

[15] M. Poesio, S. Ponzetto, and Y. Versley. Computational models of anaphora resolution: A survey. *Linguistic Issues in Language Technology*, 2011.

[16] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27:12:1–12:19, March 2009.

[17] Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. Bart: A

modular toolkit for coreference resolution. In *ACL (Demo Papers)*, pages 9–12. The Association for Computer Linguistics, 2008.

[18] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[19] I. H. Witten, Eibe Frank, and Mark A. Hall. *Data mining : practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, CA, USA, 3rd edition, 2011.

[20] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427 – 434, November 2003.