UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DETECÇÃO DE OPINIÕES E ANÁLISE DE POLARIDADE EM DOCUMENTOS FINANCEIROS COM MÚLTIPLAS ENTIDADES.

Josiane Rodrigues da Silva

Março de 2015

Manaus - AM

UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DETECÇÃO DE OPINIÕES E ANÁLISE DE POLARIDADE EM DOCUMENTOS FINANCEIROS COM MÚLTIPLAS ENTIDADES.

Josiane Rodrigues da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Computação - IComp, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientador: Marco Antônio Pinheiro de Cristo

Março de 2015

Manaus - AM

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Silva, Josiane Rodrigues da
S586d Detecção de Opiniões e Análise de Polaridade em Documentos Financeiros com Múltiplas Entidades. / Josiane Rodrigues da Silva. 2015
61 f.: il. color; 29 cm.

Orientador: Prof. Dr. Marco Antônio Pinheiro de Cristo
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Análise de Polaridade. 2. Detecção de Subjetividade. 3. Aprendizagem de Máquina. 4. Resolução de Anáfora. I. Cristo, Prof. Dr. Marco Antônio Pinheiro de II. Universidade Federal do Amazonas III. Título

DETECÇÃO DE OPINIÕES E ANÁLISE DE POLARIDADE EM DOCUMENTOS FINANCEIROS COM MÚLTIPLAS ENTIDADES.

Josiane Rodrigues da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO DO INSTITUTO DE COMPUTAÇÃO DA UNIVERSIDADE FEDERAL DO AMAZONAS COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM INFORMÁTICA.

Aprovado por:

Prof. Marco Antônio Pinheiro de Cristo, Doutor

Prof. David Braga Fernandes de Oliveira, Doutor

Prof. Thierson Couto Rosa, Doutor

MARÇO DE 2015

MANAUS, AM – BRASIL

Aos meus pais, razão de tudo em minha vida, e às minhas irmãs Joélia e Jomara.

Agradecimentos

A Deus, por permitir realizar meus sonhos, por me dar forças para superar as dificuldades e não me deixar desistir.

Minha eterna gratidão aos meus pais, que mesmo longe me acompanharam em todas as etapas da minha vida e com muito esforço e dedicação me ajudaram a concluir meus estudos. Portanto, não poderia deixar de agradecer pelo apoio e pelos incentivos que sempre me deram, principalmente nos momentos de extrema indecisão.

Às minhas irmãs queridas pelos incentivos de todos os dias, por fazerem me sentir tão amada na ausência de nossos pais.

Ao professor Marco Cristo pelos ensinamentos e pela dedicação e paciência com que sempre me orientou.

À minha amiga Lídia Lizziane, por suportar todas as minhas chatices, por sempre me fazer acreditar que no final tudo daria certo e por ser essa amiga que sempre posso contar.

Aos amigos de mestrado que me acompanharam nos momentos difíceis, mas que também compartilharam momentos alegres comigo.

Enfim, sempre corro o risco de ser injusta com alguém, mas agradeço a todos os amigos e familiares que sempre torceram por mim e que de alguma forma contribuíram para concluir mais esta etapa.

A todos vocês meu muito obrigada.

"Como é feliz o homem que acha a sabedoria, o homem que obtém entendimento, pois a sabedoria é mais proveitosa do que a prata e rende mais do que o ouro. É mais preciosa do que rubis; nada do que você possa desejar se compara a ela." (Provérbios 3:13-15)

Resumo

Análise de polaridade consiste em classificar a opinião do autor em positiva, negativa e neutra. No entanto, dado o grande volume de informações disponíveis na *Web*, esta análise manual torna-se inviável. Em particular, no domínio financeiro este tipo de análise é útil para empresas na tomada de decisões relacionadas ao mercado financeiro que parece ser particularmente propenso a mudanças de acordo com opiniões. Os trabalhos disponíveis na literatura propõem abordagens globais para esta tarefa, ou seja, consideram que o texto tem apenas uma polaridade. No entanto, verifica-se que os documentos, em sua grande maioria, citam várias entidades e as polaridades para estas entidades, em geral, são diferentes. Isto sugere que a classificação de polaridade deve ser feita em nível de entidade. Contudo, a maioria das abordagens tradicionais não concentram-se na tarefa de classificar polaridade por entidade. Além disso, observamos que muitos dos documentos no domínio financeiro nem sempre emitem opinião. Assim, uma primeira tarefa de interesse nesse domínio é identificar os documentos em que opiniões são expressas, isto é, documentos subjetivos. Portanto, neste trabalho propomos um método supervisionado para classificação de polaridade baseado em múltiplos modelos com o intuito de classificar documentos financeiros com múltiplas entidades. Em particular, estudamos estratégias de segmentação em texto que usam heurísticas de casamento de *string* e resolução de anáfora e propomos um método de classificação hierárquica baseada em detecção de subjetividade. Nossos resultados mostraram que uma abordagem baseada em múltiplos modelos é capaz de obter ganhos significativos sobre uma abordagem baseada em modelo global na tarefa de classificação de polaridade com múltiplas entidades. A segmentação do documento em sentenças que mencionam as entidades e a adoção de uma estratégia hierárquica também obtiveram ganhos, embora modestos.

PALAVRAS-CHAVE: Análise de Polaridade, Detecção de Subjetividade, Aprendizagem de Máquina, Resolução de Anáfora.

Abstract

Polarity analysis aims at classifying the author's opinion into positive, negative, or neutral. However, given the sheer volume of information available on the web, manually carrying out such task is unfeasible. In particular, in the financial domain this type of analysis is useful for companies in making decisions related to the financial market which is particularly prone to changes according to shifting of opinions. Most studies in literature deal with this problem by considering that documents have a global polarity. However, in general, documents cite several entities with possibly different polarities. This suggests that the classification should be performed in an entity level. Besides this problem, we also noted that many financial documents do not always emit opinion. Thus, a first task of interest in this research field is to identify documents on which opinions are expressed, that is, the subjective ones. Therefore, in this paper we propose a supervised polarity classification method based on multiple models to deal with financial documents with multiple entities. In particular, we study text segmentation strategies that use heuristics such as string matching and anaphora resolution and we propose a hierarchical classification method based on subjectivity detection. Our results showed that the multiple-models approach significantly outperformed the global-model baseline. The segmentation of the documents restricted to sentences that mention entities and the adoption of a hierarchical strategy also achieved gains, although modest.

KEY-WORDS: Polarity Analysis, Detection of Subjectivity, Machine Learning, Anaphora Resolution

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A internet é um grande repositório de informação não estruturada e para obter vantagem sobre estas informações é necessário compreender de alguma forma o conteúdo que é disponibilizado. No entanto, este grande volume de informações torna inviável a análise manual do conteúdo que é publicado em *web sites*, fóruns e páginas de notícias. Entre as análises que podem ser realizadas uma de crescente interesse é a *análise de polaridade*.

A análise de polaridade consiste em determinar a polaridade da opinião do autor em relação ao objeto em discussão, como por exemplo, verificar se a opinião é favorável, neutra ou negativa em relação a esse objeto. Opiniões têm grande influência sobre o comportamento das pessoas. Decisões simples como qual filme ver, qual carro comprar ou em qual ação investir eram frequentemente tomadas com base em opiniões de pessoas próximas, de especialistas ou de estudos conduzidos por instituições especializadas. Contudo, a popularidade das mídias sociais alterou este cenário, tornando acessível aos indivíduos e organizações conteúdo de opinião diversificado e em grandes volumes. Isto aumenta as opções dos indivíduos na busca de opiniões, pois não estão mais limitados à sua rede pessoal de contatos [6].

Desta forma, a análise de polaridade é usada em uma variedade de domínios de aplicação, quer seja financeiro, em ambientes online ou mesmo no domínio musical. Por exemplo, é útil para inferir automaticamente a opinião de um revisor a partir de resenhas que ele escreveu sobre um filme, a opinião de um cliente sobre um determinado produto de uma loja na internet com base no comentário que este postou, a opinião de uma pessoa sobre algo que foi postado em uma rede social etc.

Umas das áreas de interesse em análise de polaridade, e que é foco deste trabalho, é a

financeira. Em documentos financeiros é comum que opiniões sejam emitidas a respeito de empresas (entidades de interesse). Estas opiniões são importantes na medida que podem afetar o desempenho das entidades citadas em mercados de ações. Normalmente, em documentos financeiros várias entidades são citadas, o que implica que um documento possa apresentar tantas polaridades quanto as entidades que ele cita. Outro aspecto característico de documentos financeiros, em contraste com documentos de natureza puramente editorial como resenhas de livros ou filmes, é que nem sempre eles expressam opiniões ou, muitas vezes, expressam apenas opiniões neutras. Em geral, tais documentos têm menor impacto no desempenho das entidades que eles citam. Assim, uma primeira tarefa de interesse nesse domínio é identificar os documentos em que opiniões são expressas. Identificados esses documentos, uma segunda tarefa seria classificar as suas polaridades.

Amazon's new Kindle Fire was a hot itemduring the holiday shopping season, and one analyst believes the new Amazon tablet may have cost Apple well over $1 billion in holiday iPad sales. Morgan Keegan analyst Travis McCourt on Tuesday lowered his December-quarter iPad sales estimate from 16 million units to 13 million. Hot sales of the Kindle Fire ahead of the holidays are responsible for trimming sales of Apple's iPad by between 1 million and 2 million units, the analyst believes, making Amazon's new slate the main reason for McCourt's slashed forecast. On the low end of McCourt's estimate, the Kindle Fire cost Apple at least $500 million considering the iPad 2's $500 entry-level price point. If the Kindle Fire was indeed responsible for cutting iPad sales by 2 million units, Amazon tablet sales cost Apple a minimum of $1 billion. Considering the range of available iPad 2 models that sell for between $500 and $830 each, however, that gure would likely be signicantly higher. Amazon announced last week that it sold more than 4 million Kindles during the holiday shopping season, noting that the Kindle Fire was its most popular device. McCourt believes total Kindle Fire sales for the 2011 holiday shopping season were between 4 million and 4.5 million units.

Tabela 1.1: Exemplo de documento com múltiplas entidades

Na Tabela 1.1, é apresentado um documento onde são citadas duas entidades: *Amazon* e *Apple*. Observe que a polaridade para cada entidade citada é distinta. É comum em trabalhos na literatura [10][4] se inferir a polaridade do texto como um todo para apenas uma entidade pré-determinada (por exemplo, por meio de uma consulta). Em muitos destes trabalhos, um método supervisionado é aplicado para aprender um modelo que descreve a polaridade de uma entidade qualquer, dados exemplos de documentos pré-

classificados (onde é comum que o rótulo atribuído ao documento indique sua polaridade global). Ou seja, um único modelo global de polaridade é aprendido para a coleção.

No exemplo da Tabela 1.1, no entanto, a polaridade é diferente para cada entidade, uma vez que é positiva para *Amazon* e negativa para *Apple*. Logo, um mesmo documento pode ter diferentes polaridades. Além disso, um documento pode ser visto como uma combinação de fragmentos de texto relacionados a diferentes entidades e não como um todo que se relaciona com todas as entidades. No exemplo dado, enquanto os dois primeiros parágrafos se referem à *Amazon* e *Apple*, o último se refere apenas à *Amazon.*

Assim, uma estratégia a ser considerada para este problema poderia ser baseada em vários modelos, um por entidade, para descrever polaridades. O modelo de uma certa entidade poderia ser criado a partir de documentos (ou fragmentos de documentos) que citam aquela entidade. Os rótulos dos exemplos de treino deveriam ser rótulos definidos para as entidades em lugar de rótulos definidos para os documentos.

1.1 Motivação

A disponibilização de ferramentas capazes de fazer inferência de polaridade, quer seja em documentos financeiros ou de qualquer outro domínio, de forma rápida e eficiente é um grande desafio. Em particular, no domínio financeiro o grande interesse deve-se ao fato de que notícias de caráter positivo ou negativo, relacionadas a uma determinada companhia, podem afetar o desempenho desta companhia na bolsa de valores [4]. Além disso, o mercado financeiro parece ser particularmente propenso a mudanças de acordo com opiniões muitas vezes baseadas em rumores ou exageros [10].

Assim, a disponibilização de métodos automáticos para análise de polaridades em documentos financeiros é importante, pois a polaridade de um documento desta natureza poderia ser usada para ajudar a prever tendências relacionadas com o desempenho de uma companhia.

Além disso, a análise de polaridade envolvendo múltiplas entidades é de interesse para muitas aplicações, além das financeiras. De fato, qualquer texto de caráter não estritamente editorial em que se possa emitir opiniões (relacionadas com múltiplos objetos de interesse) pode ser alvo destes métodos. Este é o caso de *posts* em redes sociais sobre músicas e artistas, por exemplo, onde nem todos os *posts* emitem opiniões (alguns ape-

nas fornecem informações sobre eventos) e os que emitem podem ter como alvo vários objetos de interesse (diferentes músicas ou artistas).

1.2 Questões de Pesquisa

Neste trabalho, estudamos como melhorar a acurácia de um modelo de classificação de polaridade em documentos financeiros. Em particular, nosso estudo foi dirigido por quatro questões de pesquisa que pretendemos responder com o desenvolvimento deste trabalho. Estas questões são descritas a seguir.

- Muitos trabalhos na literatura propõem abordagens globais para o problema de análise de polaridade, ou seja, consideram que o documento tem apenas uma entidade, que é o próprio documento. Entretanto, em geral, sabemos que nestes documentos várias entidades são citadas e que as polaridades podem ser distintas para cada uma delas. Deste modo, pretendemos responder a seguinte questão: *um método baseado em múltiplos modelos (um por entidade) é melhor que um método que infere polaridade globalmente?*

- Nós também observamos que considerar todo o conteúdo do documento para inferir polaridade, para uma certa entidade, pode dificultar o aprendizado do classificador, uma vez que nem sempre todo o conteúdo se refere à esta entidade. Isso sugere que o descarte dos fragmentos não relacionados às entidades de interesse pode minimizar o ruído nos dados de treino e, assim, melhorar a classificação dos documentos em relação às entidades. Portanto, outra questão de pesquisa seria: *qual o impacto das técnicas de segmentação de texto para este problema? Existe alguma técnica em particular, como heurísticas de casamento (matching) de strings e técnicas de resolução de anáfora, que melhore o desempenho do classificador?*

- Ao estudar o problema de análise de polaridade queremos classificar um documento em relação à sua polaridade: positiva, negativa e neutra. No entanto, sabemos que muitos destes documentos não emitem opinião, ou simplesmente emitem opiniões neutras, e que classificar polaridade neutra é sempre uma tarefa difícil. Com base no problema, surge a seguinte questão: *um método que infere polaridade em múltiplos níveis (um para detectar opinião, seguido de outro para classificar polaridades*

positivas e negativas) é melhor que um método de uma única fase que classifica as três polaridades? Como este método deveria explorar a coleção de dados, em relação à distribuição das polaridades, de forma a maximizar o desempenho do classificador?

- Finalmente, verificamos que existem duas formas fundamentais de inferir polaridade: abordagem supervisionada e abordagem não supervisionada. Os métodos que utilizam abordagem supervisionada dispõem de exemplos de treino rotulados para treinar o modelo e os novos exemplos são classificados com base no que foi apresentado anteriormente, enquanto que os métodos não supervisionados classificam polaridade com base em léxicos de polaridade. Verifica-se assim que os métodos supervisionados têm vantagem sobre os não supervisionados, uma vez que estes utilizam de informação a priori. No entanto, uma desvantagem de tais métodos é que dependem de dados rotulados e em aplicações reais isso é sempre uma limitação. Isto posto, chegamos a nossa última questão de pesquisa: *quão competitivo é um método não supervisionado quando comparado a um supervisionado?*

Em nossa busca por respostas adequadas, analisamos o impacto destas questões em relação ao problema de análise de polaridade. Em particular, propomos um método que explora tais ideias, verificando sempre a melhor forma de aplicar as técnicas existentes na literatura.

1.3 Objetivos

Nesta seção, apresentamos os objetivos que visamos alcançar por meio deste trabalho de pesquisa.

1.3.1 Objetivo Geral

Definir um método automático para análise de polaridade em documentos financeiros capaz de obter resultados relevantes em uma base de dados por meio das tarefas de detecção de opinião e classificação de polaridade, explorando estratégias de segmentação em texto que citam múltiplas entidades. O modelo desenvolvimento será avaliado sobre uma

coleção de mil documentos extraídos de *sites* financeiros e rotulados para um conjunto pré-definido de entidades.

1.3.2 Objetivos Específicos

1. Selecionar e implementar estratégias de segmentação em texto presentes na literatura.

2. Definir um método para detecção de opinião que explore a subjetividade nos documentos.

3. Implementar um método baseado em múltiplos modelos, um por entidade, para a tarefa de análise de polaridade, baseados nas estratégias selecionadas e no modelo de detecção de opinião.

4. Comparar os vários modelos com os existentes na literatura com o intuito de determinar a melhor abordagem para resolver o problema de análise de polaridade, visando responder as questões apresentadas na Seção 1.2.

1.4 Organização da Dissertação

Esta dissertação está estruturada como segue. No Capítulo 2, são apresentados conceitos importantes para o entendimento do trabalho, bem como os trabalhos relacionados a análise de sentimento. No Capítulo 3, descrevemos a solução proposta para o problema apresentado. Os experimentos realizados e os resultados obtidos estão descritos no Capítulo 4 e, por fim, no Capítulo 5 discutimos as conclusões e o direcionamento para os trabalhos futuros.

Capítulo 2

Fundamentos

Este capítulo introduz conceitos básicos necessários para melhor compreensão do método proposto, bem como uma revisão da literatura relacionada à área de abrangência deste trabalho. Os conceitos apresentados incluem análise de sentimento (análise de polaridade e suas aplicações), técnicas comumente usadas para inferir polaridade, resolução de anáfora e as métricas de avaliação utilizadas.

2.1 Análise de Sentimento

A análise de sentimento é a área do conhecimento que analisa a opinião das pessoas, tais como sentimentos, avaliações, atitudes e emoções em relação às entidades. Pang [24] categoriza a área de análise de sentimentos da seguinte forma: identificação de opinião, análise de polaridade das opiniões, classificação de documentos de acordo com sua perspectiva e reconhecimento de emoção/humor. Este trabalho trata especificamente da análise de polaridade das opiniões, que segundo Liu [16], visa determinar a atitude de um agente (isto é, o dono da opinião) no que diz respeito a um determinado alvo (isto é, o receptor da opinião) que pode ser, por exemplo: um tópico, um produto, um serviço, uma entidade ou mesmo uma propriedade destes objetos. O alvo é normalmente um conteúdo de texto de um determinado contexto e tempo. Embora a atitude de um agente possa corresponder a uma decisão complexa, é comum que este seja representado por uma classificação com valores simbólicos como positivo, negativo e neutro.

2.1.1 Aplicações de Análise de Polaridade

Opiniões são centrais para quase todas as atividades humanas, porque são os principais influenciadores de nossos comportamentos. Sempre que precisamos tomar uma decisão, queremos saber a opinião dos outros. No mundo real, as empresas e as organizações sempre querem encontrar consumidores potenciais ou determinar a opinião do público sobre os seus produtos e serviços. Os consumidores individuais também querem saber as opiniões dos usuários existentes de um produto antes de comprá-lo, e as opiniões dos outros sobre candidatos políticos antes de tomar uma decisão de voto em uma eleição política.

No passado, quando alguém precisava de opiniões de qualquer natureza, a única alternativa era consultar pessoas de seu círculo de conhecimento (amigos e familiares). Quando uma organização precisava de opiniões públicas, necessariamente tinha que consultar pessoas através de pesquisas de opinião. Com o crescimento explosivo da mídia social (por exemplo, fóruns de discussão, blogs, Twitter, comentários e postagens em *sites* de redes sociais) na *Web*, indivíduos e organizações usam cada vez mais o conteúdo destes meios de comunicação para a tomada de decisão. Hoje, se alguém quer comprar um produto de consumo, esta pessoa não está mais limitada a pedir opiniões de seus amigos e familiares, porque há muitos comentários e discussões em fóruns públicos na *Web* sobre o produto. Para uma organização já não é mais necessário realizar estudos e pesquisas de opinião, a fim de recolher opiniões públicas, porque há uma abundância de tais informações publicamente disponíveis.

No entanto, encontrar e monitorar *sites* de opinião na *Web*, filtrando as informações contidas nela continua a ser uma tarefa difícil por causa da proliferação de diversos *sites*. Cada *site* geralmente contém um grande volume de texto de opinião que nem sempre é facilmente decifrado. O usuário terá dificuldade em identificar *sites* relevantes e extrair e sintetizar as opiniões deles. Os sistemas automatizados de análise de sentimento são, portanto, necessários.

Desta forma, podemos verificar inúmeras aplicações em que a análise de polaridade pode ser empregada. A seguir estão listadas algumas delas:

- *Reviews*: *sites* de comércio eletrônico, filmes, músicas 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 recomendação ajuda em não recomendar produtos com *reviews* negativos. Em sistemas de propagandas será mostrado apenas o texto que tiver um contexto positivo, ocultando dos usuário os produtos com avaliações negativas.

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

- Mercado Financeiro: uma das principais características do mercado financeiro é a avaliação dos investidores e profissionais sobre uma companhia. Estas 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 opiniões positivas e negativas, o que ajuda os investigadores na tomada de decisões.

2.2 Classificação de Polaridade

Do ponto de vista operacional, a análise de polaridade implica no uso de técnicas de análise de texto, processamento de linguagem natural e linguística computacional para identificar, extrair e entender a subjetividade do conteúdo. Na literatura, a tarefa de análise de polaridade é abordada de duas formas principais: técnicas léxicas, baseadas em dicionários, e técnicas baseadas em aprendizado de máquina. A seguir apresentamos cada uma destas técnicas.

2.2.1 Abordagem Baseada em Dicionários

Na abordagem baseada em dicionários, também conhecida como abordagem léxica ou linguística, a polaridade é inferida a partir das palavras. A Figura 2.1 ilustra o processo de classificação usando este tipo de abordagem. Como podemos observar, a fase de classificação baseia-se no uso de léxicos (dicionários) de sentimentos, que são compilações de palavras ou expressões de sentimento associadas às suas respectivas polaridades. Em geral estas técnicas são não supervisionadas e dependem de operações de pré-processamento e transformações específicas, tais como reconhecimento de *n*-gramas, extração de *features*, eliminação de termos irrelevantes, transformação do texto em vetor de termos etc.

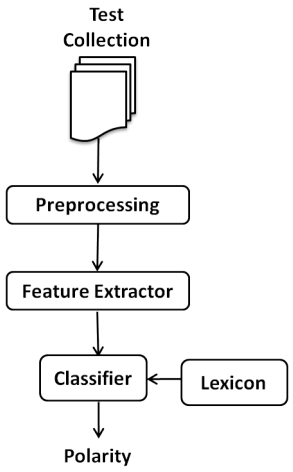

Figura 2.1: Processo de classificação de polaridade usando abordagem léxica.

Um dos métodos mais utilizados na abordagem linguística é o da co-ocorrência entre alvo e sentimento, que não leva em consideração nem a ordem dos termos dentro de um documento (*bag-of-words*) nem suas relações léxico-sintáticas. Nesta técnica, cada documento é representado por um vetor de palavras que ocorrem no documento e a polaridade é classificada segundo a polaridade individual de cada palavra. A avaliação do termo com respeito à sua classe (positiva, negativa ou neutra) é feita por meio de um dicionário que contém as informações de polaridade destes termos. Um exemplo de léxico para o caso da língua Inglesa, é a WordNet[1]. De acordo com a WordNet, a polaridade de palavras como "*good*" e "*happy*" é positiva, enquanto que para "*bad*" e "*sad*" é negativa. Desta forma, a polaridade do documento é baseada nos termos com polaridade mais frequente [24]. No entanto, a frequência do termo depende do domínio do seu documento e o seu significado, do contexto em que é usado. Assim, aplicar um conjunto de termos de um domínio em outro pode comprometer o desempenho do classificador de polaridade.

Outras técnicas são baseadas em informações linguísticas mais complexas, como a sua natureza estrutural, morfológica e sintática (rótulos de partes do discurso ou *part-of-speech tags*), ou baseadas na posição do termo no texto. É comum também identificar o relacionamento entre os termos, obtendo-os em sequência (*n*-gramas) para identificar seus significados como um todo. Por exemplo, o *n*-grama "*very good*" é interpretado como mais positivo que "*good*", enquanto que o *n*-grama "*not very good*" é interpretado como negativo devido a inversão do significado causado pelo termo "*not*". Muitas abordagens que utilizam técnicas complexas de PLN têm sido usadas para interpretar sentenças e classificar polaridades [19, 25].

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

2.2.2 Abordagem Baseada em Aprendizagem de Máquina

Além das técnicas baseadas em palavras e sentenças, usando ou não PLN, muitas abordagens são baseadas em aprendizagem de máquina.

A área de aprendizagem de máquina estuda o desenvolvimento de métodos capazes de inferir conhecimento a partir de amostras de dados. Neste sentido, alguns algoritmos são propostos no intuito de permitir que, após um determinado treinamento sobre um conjunto de dados com instâncias de classificação conhecidas, uma máquina seja capaz de interpretar novos dados e classificá-los de maneira apropriada a partir de uma generalização do que lhe foi apresentado anteriormente.

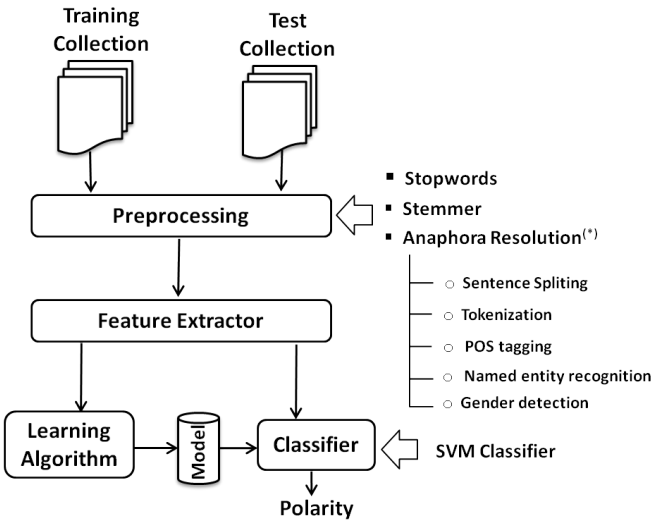

Figura 2.2: Processo classificação de polaridade usando aprendizagem de máquina.

Assim, o problema de análise de polaridade pode ser visto como uma tarefa de classificação supervisionada e muitas informações usadas por métodos anteriores, baseadas em palavras e *n*-gramas, são usadas como caraterísticas para representar os documentos a serem classificados. A Figura 2.2 exemplifica este processo: o problema de classificação é dividido em dois passos: (1) aprender um modelo de classificação sobre uma coleção de treinamento com as classes consideradas (ex.: positivo, negativo); e (2) prever a polaridade de novas porções de texto com base no modelo resultante. Dentre os algoritmos de classificação, podemos citar como os mais utilizados *Support Vector Machine*, Naive Bayes, *Maximum Entropy* e algoritmos baseados em redes neurais [6]. Como mostrado em [25], classificadores supervisionados são muito usados em trabalhos de pesquisa desta área. Uma das grandes limitações no uso de aprendizado supervisionado para definição de polaridade é a necessidade de dados rotulados para treino. O desempenho destes mé-

todos é afetado não somente pela quantidade, mas igualmente pela qualidade dos dados de treino disponíveis.

Como em Azar [4], entre os classificadores usados para classificação de polaridade [5, 25], nós adotamos *Support Vector Machine* (SVM) [5, 33]. SVM é um método para classificação binária, onde o documento o ser classificado é representado como um vetor de características. A ideia por trás do SVM é encontrar um hiperplano de separação ótima que divida duas classes de documentos, C_a e C_b. Este hiperplano é aprendido a partir do conjunto de treino de documentos anotados por humanos. A Figura 2.3 ilustra documentos de duas classes representadas por círculos (C_a) e retângulos (C_b). O SVM modela cada classe como uma região em um espaço vetorial. Os vetores marginais de cada região (também denominados vetores de suporte) são utilizados para determinar a margem de separação entre as classes. Com estes vetores, SVM encontra um vetor w, a partir do qual é gerado o hiperplano ótimo, por meio da solução do problema de otimização quadrática. Há casos em que não é possível encontrar um hiperplano que seja capaz de separar linearmente as classes, então SVM define uma margem de erro aceitável usando variáveis *slack*. Os erros associados a estas variáveis são também considerados na solução de problemas de otimização pelo SVM.

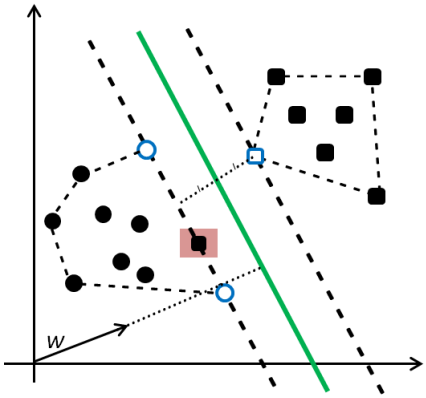

Figura 2.3: Exemplo de duas classes de documentos no espaço. A linha verde indica o hiperplano de separação gerado pelo vetor w. Os vetores de suporte estão representados pelos círculos e retângulos azuis existentes nas linhas pontilhadas. Note que um documento da classe C_b está localizado do lado esquerdo do hiperplano, o lado direito do hiperplano corresponde a C_a.

Para classificar novos documentos, esses são projetados no espaço e a classe é determinada por meio dos documentos presentes em relação ao hiperplano de separação. Observe que quanto mais próximo o documento é projetado ao hiperplano gerado por w, mais difícil é classificá-lo. Assim, documentos projetados mais distantes do hiperplano

são classificados com mais confiança.

Para usar SVM para classificar mais de duas classes, muitas abordagens são propostas. Uma estratégia comum é reduzir o problema de classificação em *n* problemas de classificação binária [5]. No final, decisões binárias são combinadas em uma decisão final. Para classificar documentos com SVM, nós utilizamos a ferramenta LIBSVM[2] [8]

2.3 Resolução de Anáfora

Uma estratégia simples para determinar se uma entidade é citada em uma sentença é verificando a ocorrência da *string* correspondente ao nome da entidade nos *n*-gramas extraídos da sentença. Por exemplo, considere a sentença "*Cats are very clean*" e "*However, they always get dirty when they go outside*". É possível inferir que a primeira sentença se refere a "*Cats*" pela simples verificação da ocorrência da *string* "*Cats*". No entanto, esta abordagem não funciona para a segunda sentença, em que a entidade "*Cats* foi representada pelo pronome "*they*". O problema de determinar que fragmentos de texto referem-se a uma mesma entidade é chamada de resolução de anáfora ou resolução de co-referência. Este problema deve-se ao fato de que uma mesma entidade pode ser referenciada por diferentes expressões linguísticas. Nas sentenças anteriores, "*Cats*" e "*they*" se referem à entidade "*Cats*".

Muitos estudos em processamento de linguagem natural [27, 22] têm abordado o problema de resolução de anáfora. Mais formalmente, o problema de anáfora pode ser definido como: um substantivo A é um antecedente anafórico de B se e somente se A é necessário para interpretação de B.

Para resolver anáfora, o texto é primeiramente segmentado em sentenças usando um separador de sentença. Em seguida, os elementos da sentença (tal como os nomes das entidades, substantivos, verbos e advérbios) são identificados. A identificação dos nomes das entidades e substantivos é essencial para resolver anáfora, uma vez que eles são comumente usados para descrever pessoas, lugares, objetos e conceitos. Esta identificação requer conhecimento da gramática da língua alvo, que descrevem padrões de uso e informações de domínio específico. Por exemplo, em Inglês, os substantivos podem ser identificados através do reconhecimento de outras estruturas linguísticas, tais como pro-

[2]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

nomes definidos (ex.,"*Ross bought {a MP3 player / three flowers} and gave {it / them} to Nadia for her birthday*"), pronomes indefinidos (ex.: *one* em "*Kim bought a t-shirt so Robin decided to buy one as well*"), pronomes demonstrativos (ex.: "*that*"), nominais (ex.: "*a man*", "*a woman*" e "*the man*") e nomes próprios (ex.: "*John*", "*Mary*").

Outros elementos da sentença, tais como verbos, adjetivos, preposições e advébios, são úteis para definir os lugares onde os substantivos devem aparecer na frase. O conhecimento destes elementos de sentença são fornecidos por rotuladores de *part-of-speech* (POS tagger) [27, 22], juntamente com um analisador morfológico (*shallow parser*) capaz de reconhecer elementos simples e compostos não considerando sua estrutura interna. Já o *POS tagger* identifica a categoria morfossintática de um *token*. Assim, em geral, uma coleção rotulada é usada para treinar um algoritmo de aprendizado. Este algoritmo também usa o contexto em que o *token* é usado para determinar sua categoria morfossintática.

Neste trabalho, nós exploramos resolução de anáfora usando a ferramenta *Stanford CoreNLP*[3] [18]. CoreNLP fornece um conjunto de ferramentas para análise de linguagem natural, que integra a maior parte das etapas de processamento de linguagem natural, incluindo: *part-of-speech (POS) tagger*, reconhecedor de entidades nomeadas (NER), analisador de linguagem natural que analisa a estrutura gramatical das sentenças, sistema de resolução de co-referências, análise de sentimento e ferramentas de aprendizagem de padrões fracamente supervisionadas baseadas em semente (*bootstrapping*). A distribuição básica da ferramenta fornece suporte para análise de textos em Inglês, no entanto, é possível aplicar a ferramenta em outros idiomas, tais como Chinês e Espanhol.

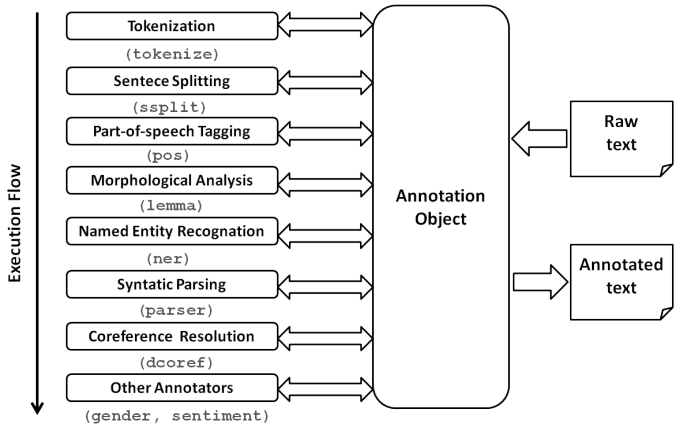

Figura 2.4: CoreNLP: Arquitetura do Sistema.

A arquitetura do sistema CoreNLP inclui todos os módulos necessários para resolver

[3]http://nlp.stanford.edu/software/corenlp.shtml#About

co-referências, como mostra a Figura 2.4. O processamento de um texto é feito seguindo as etapas:

- O texto de entrada é submetido a um processo de anotação, que consiste em uma sequência de anotadores.
- Neste processo de anotação, o texto é representado por uma sequência de *tokens*, que em seguida são agrupados em sentenças. Os *tokens* são rotulados com suas parte do discurso, são gerados os lemas, é feito o reconhecimento das entidades (se são nomes de empresas, pessoas, lugares etc.) e é fornecida uma análise sintática completa, incluindo uma representação de dependências baseada em análise probabilística. Com base nestas informações, é possível fazer análise de sentimento aplicando um modelo composicional baseado em um classificador e implementar detecção de menções e resolução de co-referências.
- Na saída é obtida uma anotação contendo todas as informações analisadas pelos anotadores, estruturadas em um arquivo XML.

2.4 Métricas de Avaliação

Métodos de classificação de polaridade são avaliados em termos de eficácia com o objetivo de medir a capacidade do classificador de classificar corretamente as amostras. Os resultados dos experimentos realizados neste trabalho foram obtidos por meio da métrica de acurácia [33, 5, 17]. A acurácia é uma métrica bastante utilizada, cujo cálculo consiste na razão entre o total de documentos corretamente classificados e o número total de documentos na coleção. Portanto, a acurácia verifica se os documentos, para os quais já se conhece as classes a qual pertencem, foram classificados corretamente.

Para a estimar a acurácia final utilizamos o teste de validação cruzada de k-$fold$ com $k = 5$ [17]. Na validação cruzada, o conjunto de dados $\mathcal{D}$ é dividido aleatoriamente em k subconjuntos mutuamente exclusivos $\mathcal{D}_1$, $\mathcal{D}_2$, ..., $\mathcal{D}_k$ de tamanhos aproximadamente iguais. O classificador é treinado e testado k vezes. Em cada iteração $i \in \{1,2,k\}$ o classificador é treinado em $\mathcal{D} \setminus \mathcal{D}_i$ e testado em $\mathcal{D}_i$. A acurácia de cada $fold$ é obtida a partir do número de classificações corretas, dividido pelo número de instâncias no conjunto de

dados. Formalmente, considere $\mathcal{D}_i$ o conjunto de teste que inclui as instâncias $x_i = \langle v_i, y_i \rangle$, a acurácia final obtida pela validação cruzada é definida pela Equação 2.1.

$$ac_{=} \frac{1}{n} \sum_{\langle v_i, y_i \rangle \in D} \delta(\Upsilon(D \setminus D_{(i)}, v_i), y_i) \tag{2.1}$$

O objetivo de usar validação cruzada é validar o modelo sobre um conjunto de dados diferente daquele usado para construção do modelo.

2.5 Trabalhos Relacionados

A análise de sentimentos tem sido empregada para solucionar diversos problemas. Embora métodos de análise possam ser baseados puramente em técnicas léxicas (onde a polaridade do texto é inferida a partir da polaridade intrínseca da palavra), métodos baseados em conjuntos mais ricos de atributos que capturam léxicos, contexto e outros elementos do texto são mais comumente usados. A seguir são apresentados alguns trabalhos da literatura que exploram o problema de análise de polaridade.

2.5.1 Métodos Baseados em Abordagens Globais

Um dos primeiros estudos em análise de polaridade foi proposto por Pang *et al.* [26], que aplicou a classificação de polaridade em uma abordagem similar a de classificação de tópicos. Usando resenhas de filmes extraídas do *site Internet Movie Database* (IMdB), os autores avaliaram três classificadores (Support Vector Machine, Máxima Entropia e Naive Bayes) para esta tarefa. Eles notaram que os classificadores, com exceção do Naive Bayes, tiveram desempenho comparável ao realizado por pessoas. Eles também observaram que os erros apresentados pelos classificadores estavam relacionados com comentários irônicos e questões de contextualização.

Wilson *et al.* [32] exploraram o problema de contextualização. Eles estudaram o problema baseando-se em conjuntos léxicos pré-classificados em positivos ou negativos, e mostraram que a polaridade depende do contexto em que a palavra é usada. Por exemplo, apesar da palavra "*trust*" expressar um contexto positivo, isso não é verificado na sentença

"*Philip Clapp is the president of the National Environment Trust*". Neste contexto, a palavra é neutra. A solução proposta pelos autores usou um método que aplica técnicas de aprendizado de máquina que explora características das sentenças. Yi *et al.* [34] também propuseram que a análise de polaridade deve ser feita em nível de sentença e sugeriram um método de classificação de polaridade baseado em conceitos de Processamento de Linguagem Natural (PLN).

O primeiro estudo em análise de polaridade no domínio financeiro foi proposto por Azar [4], motivado pela possibilidade de prever reações no mercado de ações. Ao usar heurísticas de PLN juntamente com os classificadores SVM e Árvore de Decisão, ele alcançou desempenho similar aos de anotadores humanos. No estudo feito, foram utilizados documentos citando companhias com mais de 20 notícias da base de dados *Reuters Key Developments Corpus*. Ele também observou que os modelos aprendidos para o domínio financeiro apresentam baixo desempenho em diferentes domínios.

Bollen *et al.* [7] estudaram a correlação entre a polaridade das opiniões relacionadas às companhias e seu desempenho no mercado de ações. As informações sobre as companhias foram obtidas do Twitter e foi utilizada uma simples estratégia baseada no *Google-Profile of Mood States* (GPOMS). Depois de analisar o grande volume de dados no Twitter, eles descobriram que as mudanças detectadas no estado emocional a partir de *tweets* estão correlacionadas com as mudanças observadas no mercado de ações. Muitos trabalhos subsequentes têm focado também no Twitter, como o estudo desenvolvido por Montejo-Ráez *et al.* [20], que usou uma abordagem léxica baseada na Wordnet e uma comparação de métodos apresentados em [3].

Outros trabalhos também abordaram a correlação entre mudanças emocionais e a oscilação das ações, como o estudo apresentado em [10] e [11]. Nestes trabalhos é explorada a teoria de Darwin sobre as emoções básicas do homem (raiva, medo, tristeza, alegria etc). Eles trataram as emoções como conceitos bidimensionais em vez de categorias discretas. Desta forma, cada emoção foi caracterizada de acordo com sua natureza (do mal para o bem) e intensidade (de fraca a forte). Para a avaliação da polaridade os autores construíram um grafo G representando todo o conteúdo, onde os nós em G representam palavras

com seus respectivos valores de polaridade (extraídos da Wordnet usando a ferramenta SentiWordNet). Por meio da iteração no grafo, eles foram capazes de estimar o sentimento global e sua intensidade. Para avaliar a abordagem proposta, os autores utilizaram notícias e informações de ações de duas empresas aéreas da Irlanda. O método proposto foi capaz de capturar a maior parte dos termos positivos, mas com baixa precisão. Quanto aos aspectos negativos, a revocação foi baixa, mas a precisão foi alta.

Schumaker e Chen [30] também estudaram o problema de classificação de polaridade no domínio financeiro. Os autores estudaram duas representações textuais: *bag-of-words* e frases nominais com nomes de entidades. Neste trabalho, os autores também observaram uma correlação entre os preços futuros das ações de companhias e a polaridade de notícias relacionadas a essas companhias. Outra abordagem baseada em *bag-of-words* no domínio financeiro foi proposto por Im *et al.* [14], onde o foco do trabalho foi o uso de *stemming*. Os autores observaram que *stemming* é útil para melhorar o reconhecimento de sentimento.

2.5.2 Métodos Baseados em Múltiplas Entidades

Toda a pesquisa anteriormente apresentada tratou a polaridade como um conceito relacionado com a opinião geral expressa em documentos. Enquanto que isso é comum em documentos editoriais em que as opiniões estão relacionadas a um tema central, produto ou serviço, este não é o caso de documentos não estruturados que expressam opiniões sobre várias entidades.

Entre as poucas pesquisas que exploram análise de polaridade em documentos com múltiplas entidades, temos as que se baseiam em estratégias conhecidas como composicionais. Nestas estratégias, a polaridade é estimada dentro dos sub-contextos e a polaridade global é obtida na forma de uma composição das polaridades destes sub-contextos, usando uma gramática de dependência. Note-se que, uma vez que cada contexto atômico está associado à polaridade de uma entidade, este método naturalmente infere polaridades para várias entidades. Nesta abordagem, a classificação de polaridade é quebrada em várias combinações binárias, onde dois elementos sintáticos, na entrada, são combinados de

cada vez, associados a uma polaridade lógica de três valores (positivos, negativos, neutro), que são controlados por uma gramática de sentimentos que calcula a polaridade dos elementos compostos resultantes. O processo inicia com uma análise léxica a nível de palavra, procede recursivamente via níveis sintáticos intermediários e termina em nível de sentença. Esta estratégia foi proposta pela primeira vez por Moilanen e Pulman [19]. Depois de avaliar, os autores concluíram que o método proposto apresenta acurácia ligeiramente inferior ao de anotadores humanos.

Mais tarde [13], os mesmos autores aplicaram suas técnicas para inferir polaridade no domínio político em mensagens de blog. Uma abordagem similar foi proposta por Romanyshyn [29], em que um sistema baseado em regras foi utilizado para detectar sentimento de cláusulas individuais em *reviews* para a língua Ucraniana. A composição de cláusulas de sentimentos permite uma análise que é multi-entidade. Finalmente, Ward *et al.* [31] propôs um framework para análise de sentimento em nível de entidade. A principal regra da análise é que uma entidade específica é utilizada como consulta para um conjunto de dados. O sentimento da entidade é obtido a partir do conjunto de informações no conjunto de dados retornados como respostas já analisadas.

2.5.3 Métodos Baseados em Detecção de Subjetividade

No nosso trabalho também estudamos uma abordagem de classificação hierárquica para a tarefa de análise de polaridade. Fomos motivados pelo fato de que esta tarefa pode ser vista como duas classificações distintas: (i) detecção de casos subjetivos e (ii) classificação de polaridade positiva e negativa dos casos subjetivos. Neste tipo de abordagem o algoritmo de aprendizado induz um modelo que captura os relacionamentos mais relevantes entre as classes funcionais no conjunto de dados de treinamento, considerando os relacionamentos hierárquicos entre as classes. Uma abordagem similar é usada no trabalho de Pang e Lee [23], em que a classificação de polaridade é melhorada pela remoção de sentenças objetivas do conjunto de treino. Naquele trabalho, os autores estão interessados em classificar polaridade em *reviews* de filmes. Para isso, eles primeiro aplicam um detector de subjetividade que determina se cada sentença é subjetiva ou não, descartando

as objetivas e criando uma extração que deve representar melhor o conteúdo dos *reviews* dos filmes, o que melhora o desempenho do classificador de polaridade que foi usado.

Para a classificação de subjetividade para a língua Árabe, Abdul-Mageed *et al.* [1] realizaram classificação binária em nível de sentença. Eles usaram uma coleção de dados manualmente anotada e construíram um léxico para língua Árabe composto por 3982 adjetivos, que foi elaborado a partir de artigos de notícias (extraídos do *Penn Arabic tree bank*). Eles usaram recursos que são semelhantes aos desenvolvidos por Wilson *et al.* [32]. Em um trabalho posterior, Abdul-Mageed *et al.* [2] estenderam seu trabalho para conteúdos de mídia social, incluindo sessões de chat, *tweets*, páginas de discussão da Wikipedia e fóruns *on-line*. Eles também exploraram características específicas da língua que incluem *stemming*, *POS tagging* e dialeto vs. Árabe Moderno. Ao final, os autores concluíram que *POS tagging* melhora a classificação e verificaram que a maioria dos *tweets* são subjetivos.

Outros trabalhos também exploraram o conceito de subjetividade. Em [21] os autores propõem uma abordagem para classificação de polaridade usando detecção de subjetividade em microblogs e Rafaee e Verena [28] exploraram os mesmos conceitos para o Twitter. Diferente de todos estes trabalhos, verificamos que enquanto o detector de subjetividade na primeira fase é mais eficaz em determinar os casos neutros, a sua taxa de erros ainda é alta o suficiente para não possibilitar a mesma eficácia do classificador de polaridade binária na segunda fase. Assim, um método semi-hierárquico parece mais adequado.

2.5.4 Resumo dos Trabalhos Relacionados

Para obter uma visão geral de como cada um dos trabalhos relacionados trata o problema de classificação de polaridade, os listamos na Tabela 2.1 e os classificamos em relação às ideias que são centrais para este problema. Nesta tabela, os métodos são descritos de acordo com (a) como organizam o conteúdo textual, ou seja, o que tratam como *documento*: o documento em si ou fragmentos destes documentos como sentenças; (b) a abordagem usada para classificação é supervisionada ou não supervisionada; (c) a pola-

ridade é inferida para todas as entidades citadas no texto, um sub-conjunto delas ou de forma única (globalmente para o documento, como se houvesse apenas uma entidade); e, finalmente, (d) o alvo da polaridade é a entidade, a página/documento ou o fragmento/-sentença.

Nesta tabela, o nosso trabalho é o último listado. Como podemos observar, poucos trabalhos propõem abordagens que consideram múltiplas entidades, tendo a maioria deles como alvo de polaridade o documento ou as suas sentenças. Os métodos que estamos propondo são os únicos supervisionados que tratam o problema de múltiplas entidades.

Trabalhos	**Documento**	**Sup. X Não Sup.**	***N* Entidades**	**Alvo da Polaridade**
Pang *et al*. Thumbs up?: sentiment classification using machine learning techniques. In EMNLP, 2002.	*Review*	Sup.	1	*Review*
Wilson *et al*. Recognizing contextual polarity in phrase-level sentiment analysis. HLT, 2005.	Sentença	Sup.	1	Sentença
Yi *et al*. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. ICDM, 2003.	Sentença	Não Sup.	1	Sentença
Azar, Sentiment Analysis in financial news, These Harvard, 2009.	Página	Sup.	1	Página
Bollen *et al*. Twitter mood predicts the stock market. Journal of Computational Science, 2011.	*Tweet*	Não Sup.	1	*Tweet*
Montejo-Ráez *et al*. Ranked wordnet graph for sentiment polarity classication in twitter. CSL, 2014.	*Tweet*	Não Sup.	1	*Tweet*
Araújo *et al*. Measuring sentiments in online social networks. WebMedia, 2013.	*Post*	Não Sup.	1	*Post*
Devitt and Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. AMACL, 2007.	Página	Não Sup.	1	Página
Devitt and Ahmad. A lexicon for polarity: Affective content in financial news text. LSP, 2007.	Página	Não Sup.	1	Página
Schumaker and Chen,. Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Trans. Inf. Syst, 2009.	Página	Sup.	1	Página
Im *et al*. Analysing market sentiment in financial news using lexical approach. ICOS, 2013	Página	Sup.	1	Página

Trabalhos	**Documento**	**Sup. X Não Sup.**	***N* Entidades**	**Alvo da Polaridade**
Moilanen and Pulman, Multi-entity sentiment scoring, RANLP, 2009.	Sentença	Não Sup.	*n*	Entidade/ Página
Romanyshyn. Rule-based sentiment analysis of ukrainian reviews. International Journal of Artificial Intelligence & Applications, 2013.	Sentença	Não Sup.	*n*	Entidade/ Página
Ward, Empath: A framework for evaluating entity-level sentiment analysis, CEWIT, 2011.	Página	Não Sup.	*n*	Página
Gryc and Moilanen. Leveraging textual sentiment analysis with social network modelling. In Proc. of the "From Text to Political Positions"Workshop, 2010.	Sentença	Não Sup.	*n*	Entidade/ Página
Pang and Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. ACL, 2004.	Sentença	Sup.	1	Sentença
Abdul-Mageed *et al.* Subjectivity and sentiment analysis of modern standard arabic. ACL, 2011.	Sentença	Sup.	1	Sentença
Abdul-Mageed *et al.* Samar: A system for subjectivity and sentiment analysis of arabic social media. ACL, 2012.	Sentença	Sup.	1	Sentença
Mourad and Kareem. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. CAS, 2013.	Sentença	Sup.	1	Sentença
Nossos Métodos	Sentença/ Página	Sup.	*n* (5)	Entidade/ Página

Tabela 2.1: Resumo dos Trabalhos Relacionados

Capítulo 3

Análise de Polaridade com Múltiplas Entidades

Neste capítulo, apresentamos o método baseado em múltiplos modelos de aprendizagem. Descrevemos as estratégias de segmentação de texto que foram exploradas neste trabalho, com o intuito de melhorar a classificação de polaridade em documentos com múltiplas entidades. E por fim, apresentamos o modelo de classificação baseado em detecção de subjetividade.

3.1 Aprendizado de Múltiplas Polaridades com Múltiplos Modelos de Aprendizagem

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 análise de polaridade pode ser vista como uma tarefa de classificação onde, para cada documento d_i, o objetivo é prever o rótulo y_i, ou seja, encontrar a função (modelo) $f : \mathcal{D} \implies \{+, -, N\}$, tal que $f(d_i) = y_i$.

No problema de análise de polaridade com m entidades $\{E_1, E_2, ..., E_m\}$, cada documento d_i é associado com um conjunto de polaridades $Y_i = \{y_{i1}, y_{i2}, ..., y_{im}\}$. Deste modo, a polaridade y_{ij} corresponde a polaridade de d_i com respeito a entidade E_j. Neste caso, a tarefa de análise de polaridade pode ser vista como uma classificação onde, dada uma

coleção de documentos $\mathcal{D}$, o objetivo é encontrar m funções $f_j : \mathcal{D} \implies \{+,-,N\}$, tal que $f_j(d_i) = y_{ij}$, $1 \leq j \leq m$.

Note que documentos onde a entidade E_j não ocorre, provavelmente não contribui para o aprendizado da função f_j, visto que dificilmente poderia ser considerado bons exemplos de casos positivos, negativos ou neutros para E_j. Portando, a função f_j deve ser melhor representada como $f_j : \mathcal{D}_j \implies \{+,-,N\}$, onde D_j é o subconjunto de documentos de $\mathcal{D}$ que cita a entidade E_j.

Do mesmo modo, sentenças de documento que citam uma única entidade E_j não devem ser usadas como exemplos de treino para a entidade E_k, $k \neq j$. Suponha que uma sentença (ou o documento inteiro) s é avaliada como negativa. Pode não ser adequado usar s como um exemplo negativo para E_k se s não cita E_k. Por exemplo, no documento mostrado na Figura 3.1, o primeiro parágrafo cita apenas Nokia. Assim, não está claro que o elemento poderia ser um bom exemplo de treino para entidades tais como Apple, Google e Microsoft, mesmo que estas três companhias sejam citadas nos outros parágrafos do documento.

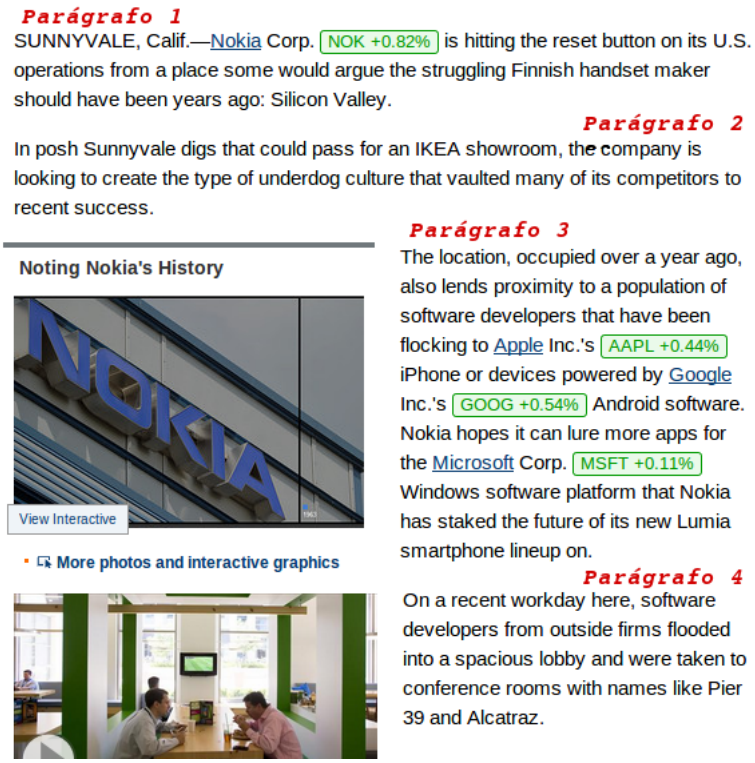

Figura 3.1: Documento com citação para Nokia, Apple, Microsoft e Google.

Desta forma, considere d_i^j um documento composto por todas as sentenças de d_i que

cita E_j. Notamos por $\mathcal{D}^{(j)}$ o conjunto de todos os documentos d_i^j, ou seja, o conjunto de todos os documentos que são compostos apenas pelas sentenças que citam E_j. Dadas as definições, podemos reescrever f_j como $f_j : \mathcal{D}^{(j)} \implies \{+,-,N\}$.

Podemos agora definir duas estratégias para aprender polaridade para múltiplas entidades usando múltiplos modelos:

- Modelo baseado em documento (DbM): onde cada função f_j é associada à coleção de documentos $\mathcal{D}_j$, ou seja, o conjunto de *documentos* que citam E_j.
- Modelo baseado em Sentença (SSM): onde cada função f_j é associada à coleção de documentos $\mathcal{D}^{(j)}$, ou seja, o conjunto de documentos compostos apenas pelas *sentenças* que citam E_j.

Portanto, verificamos que existem duas formas fundamentais de inferir polaridade: por documento e por sentença.

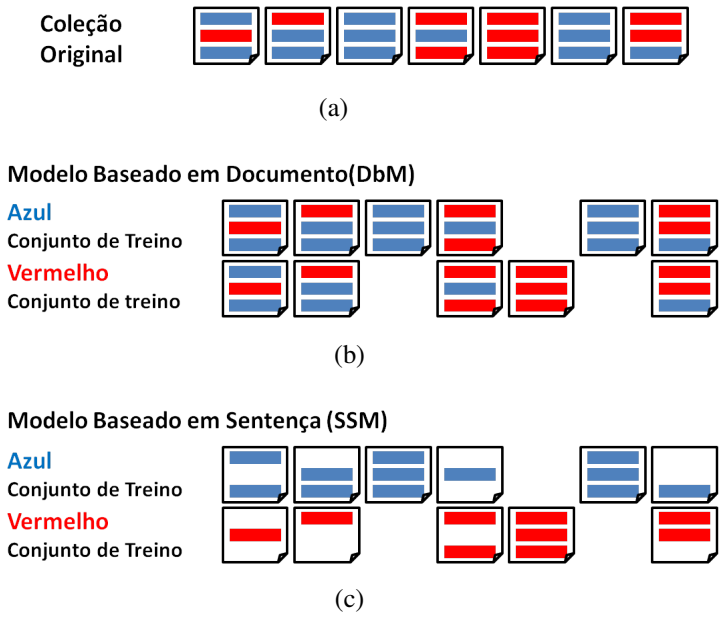

Figura 3.2: Separação do conjunto de treino para o modelo baseado em documento e em sentença.

Para entender melhor como cada uma das abordagens funcionam, observe a Figura 3.2. Consideremos inicialmente uma base de dados (Figura 3.2(a)) formada por documentos que citam duas entidades específicas: entidade Vermelho e entidade Azul. Para o modelo baseado em documento (Figura 3.2(b)), o conjunto de treino para a entidade

Azul será formado por todos os documentos que citam a entidade Azul, assim como para entidade Vermelho o modelo será treinado com todos os documentos que citam a entidade Vermelho. Observe que neste modelo é considerado o documento todo e não parte dele. Já no modelo baseado em sentença (Figura 3.2(c)), o conjunto de treino para a entidade Azul será formado apenas pelas sentenças que citam a entidade Azul, a mesma ideia seguimos para a entidade Vermelho. Desta forma, na primeira abordagem consideramos o documento inteiro e na segunda abordagem apenas as partes do documento que citam cada entidade.

Como documentos financeiros, em geral, citam várias entidades, neste caso precisamos considerar um modelo que aprende polaridade baseado em informações que de fato se referem às entidades, já que os várias fragmentos de texto presentes no documento podem se referir à diferentes entidades. Isto significa que temos que fazer esta análise visando descobrir como dividir este documento de tal forma a refletir melhor a ideia de que existem diferentes modelos, um por entidade, além da ideia global. Com intuito de explorar tais ideias, na próxima seção definimos algumas estratégias de segmentação em texto.

3.2 Mapeando Sentenças e Entidades

Como descrito anteriormente, a Figura 3.1 mostra um documento que cita as entidades Nokia, Apple, Google e Microsoft. Como podemos ver, o primeiro parágrafo se refere à Nokia, e o terceiro se refere à todas as quatro entidades, uma vez que estas entidades são diretamente citadas pelos parágrafos. Dos dois parágrafos restantes, o segundo se refere a três das quatro entidades citadas, enquanto que o último não cita nenhuma delas.

No segundo parágrafo, as entidades são citadas indiretamente. O termo "*the company*" é usado para se referir à Nokia, enquanto que o termo "*competitors*" corresponde à Apple e Google. Este tipo de citação indireta (anáfora) não é tão simples de capturar, uma vez que requer uma melhor compreensão do contexto e, as vezes, conhecimento do domínio e contexto. Compreensão do texto é entender que "*the company*" se refere a companhia que é o foco do discurso na sentença onde "*company*" é usado. Conhecimento do domínio e

contexto é saber que no tempo em que o texto foi escrito, a Nokia estava competindo com a Apple e Google, mas não com a Microsoft.

Desta forma, a tarefa de determinar que entidades uma sentença se refere pode ser realizada por métodos que variam de simples heurísticas de ocorrência de *string* à complexas abordagens baseadas em PLN, tal como como resolução de anáfora. Simples heurísticas de ocorrência de *string* aplicam-se à primeira sentença, que cita Nokia, já que a *string* "nokia" ocorre no texto. Resolução de anáfora aplica-se ao segundo parágrafo, que cita Nokia, visto que "*company*" se refere à Nokia.

Baseado nestas observações, propomos seis variantes para a estratégia SSM: três delas baseadas em heurísticas de ocorrência de *string* e outras três baseadas em resolução de anáfora. As seis variantes são descritas a seguir:

- SSM1: a sentença *s* é atribuída às entidades cujo os nomes ocorrem em *s*. Se nenhuma entidade está presente em *s*, *s* é atribuída à todas as entidades. Esta heurística tenta capturar situações como a observada no segundo parágrafo do texto da Figura 3.1;
- SSM2: a sentença *s* é atribuída às entidades cujo os nomes ocorrem em *s*. Se nenhuma entidade está presente em *s*, *s* é descartada. Esta heurística tenta capturar situações como observadas no quarto parágrafo do texto da Figura 3.1;
- SSM3: a sentença *s* é atribuída à última entidade citada se nenhuma entidade estiver presente em *s*. A intenção por trás desta heurística é que se nenhuma nova citação foi feita, o texto provavelmente ainda se refere à última entidade citada;
- SSM4: a sentença *s* é atribuída às entidades referenciadas em *s*. Se nenhuma entidade é referenciada por *s*, *s* é atribuída à todas as entidades. Esta heurística é equivalente a SSM1, porém usando resolução de anáfora;
- SSM5: a sentença *s* é atribuída à todas as entidades referenciadas em *s*. Se nenhuma entidade é referenciada por *s*, *s* é descartada. Esta heurística é equivalente a SSM2, porém usando resolução de anáfora;
- SSM6: a sentença *s* é atribuída à última entidade referenciada se nenhuma entidade

é referenciada por *s*. Esta heurística é equivalente a SSM3, porém usando resolução de anáfora;

Nas estratégias SSM1 a SSM3, a noção de citação corresponde à ocorrência do nome da entidade na sentença como, por exemplo, no caso da Nokia no primeiro parágrafo da Figura 3.1 (neste exemplo, parágrafos desempenham papel de sentença). Neste caso, o texto é primeiro pré-processado usando algum algoritmo morfológico de PLN. Em seguida, o texto é segmentado em sentenças, ou seja, a sequência de *tokens* termina com um período. *Tokens* são sequências de caracteres que correspondem às palavras. Além disso, todos *stopwords* (preposições, artigos, numerais etc) são removidos. Finalmente, os nomes das entidades são verificados se ocorrem no texto resultante.

Para ilustrar a saída dos métodos SSM1 a SSM6, considere o texto da Figura 3.1 que contém quatro parágrafos. Depois de executar SSM1, podemos verificar que o único parágrafo não atribuído à Google, Apple e Microsoft é o primeiro porque, de acordo com a definição de SSM1, Nokia é a única entidade citada nele. O documento gerado por SSM1 é quase o mesmo que o original. Deste modo, para Nokia, os parágrafos 1, 2, 3 e 4 são os fragmentos do documentos que se referem a ela. E os parágrafos 2, 3 e 4 são as sentenças que se referem à Apple, Google e Microsoft. Note que o parágrafo 1 é descartado para essas entidades, uma vez que cita apenas a Nokia.

Quanto a SSM2, a diferença consiste em descartar o parágrafo onde a entidade não está presente. Como resultado, apenas os parágrafos 1 e 3 serão considerados para a Nokia. Em relação às outras entidades, o documento será composto apenas pelo parágrafo 3.

No resultado usando SSM3, o parágrafo é atribuído para última entidade citada. Consequentemente, o parágrafo 2 é atribuído para Nokia e o parágrafo 4 para Microsoft. Nokia também é citada nos parágrafos 1 e 3, enquanto que as outras entidades são citadas apenas no parágrafo 3.

Quanto às estratégias SSM4 a SSM6, é considerado que o parágrafo cita uma entidade quando se refere a ela direta ou indiretamente. É o caso da Nokia no segundo parágrafo da Figura 3.1. Para estes casos, CoreNLP é usado para resolver as anáforas encontradas

no texto.

3.3 Detectando Subjetividade

Em análise de sentimento, muitos estudos focam na classificação de polaridade como uma tarefa que visa classificar informações em positivas, negativas e neutras, ou seja, as classes predefinidas são tratadas isoladamente e não há nenhuma estrutura que define as relações entre elas. No entanto, para muitas aplicações em PLN, a habilidade de detectar e classificar opiniões e fatos em texto oferece vantagens distintas ao decidir que informações são relevantes ou devem ser consideradas na resolução de um determinado problema. Por exemplo, aplicativos de extração de informação podem ter como alvo declarações factuais em vez de opiniões subjetivas [35].

No caso de análise de sentimento, estamos interessados na opinião do autor em relação a um objeto, então uma forma de melhorar a classificação de polaridade é remover da coleção de treino os documentos que não contêm opinião. Alguns trabalhos na literatura fornecem métodos para análise de polaridade, determinando se um documento tem ou não opinião [23, 1, 2], mas nenhum deles combina estas ideias para uma análise em nível de entidade. Desta forma, neste trabalho nós propomos um método que combina as estratégias de segmentação descritas anteriormente com um modelo de detecção de subjetividade (opinião), que é empregado em uma fase anterior a classificação de polaridade. Portanto, podemos resolver o problema de análise de polaridade desenvolvendo as tarefas a seguir.

Dado um documento d e uma entidade E, duas tarefas são realizadas:

- Classificação subjetiva: Determinar se d é um documento subjetivo ou um documento objetivo em relação a E.

- Classificação de polaridade: Se d é subjetivo em relação a E, determinar se este expressa uma opinião positiva ou negativa.

A Figura 3.3 ilustra este processo. Os documentos já segmentados são submetidos a um classificador que tentará aprender quais documentos são subjetivos, e estes por sua vez

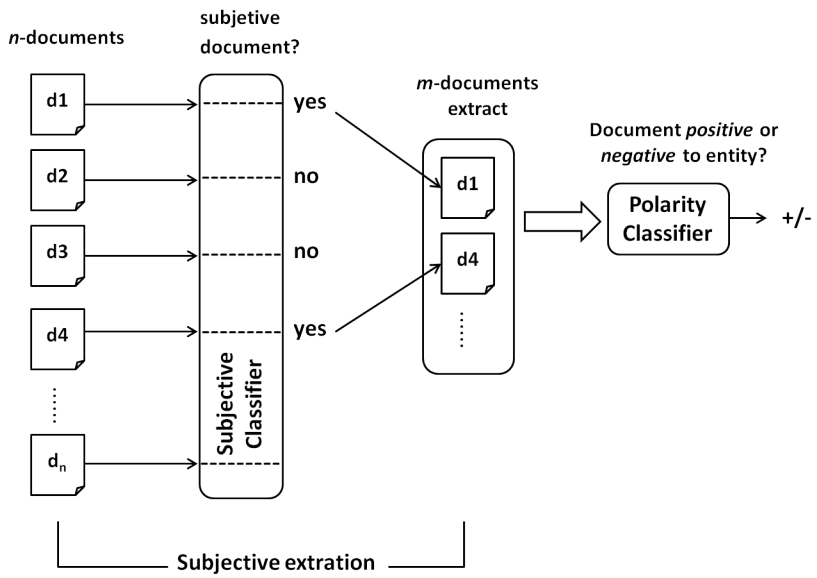

Figura 3.3: Classificação de polaridade via detecção de opinião.

são classificado quanto à sua polaridade. Neste sentindo, as duas sub-tarefas de classificação são importantes porque (1) filtra os documentos que não contêm nenhuma opinião em relação à entidade em questão e (2) uma vez filtrados estes documentos, fica mais fácil determinar se a opinião expressa neles e se suas características são positivas ou negativas. Isto pode impedir o classificador de polaridade de considerar textos irrelevantes ou mesmo enganosos: por exemplo, embora a sentença "*Samsung tries to keep its good performance*" contenha a palavra "*good*", isto não nos diz nada sobre a opinião do autor e de fato não poderia ser incorporada a uma classificação positiva sobre a empresa.

Finalmente, note que ao contrário de outros trabalhos na literatura, nós vamos adotar uma abordagem mais flexível em relação à segunda etapa da classificação, avaliando a possibilidade de não se restringir a uma taxonomia mais limitada de acordo com a taxa de erro observada na primeira etapa.

Capítulo 4

Experimentos

Neste capítulo, avaliamos o método proposto no Capítulo 3. Apresentamos a coleção sobre a qual realizamos os experimentos e mostramos os critérios para a escolha do *baseline*. Por fim, expomos e discutimos os resultados obtidos a partir da comparação entre o *baseline* e a solução proposta para o problema.

4.1 Coleção de Dados

Para avaliar o método desenvolvido, nós utilizamos uma coleção de dados composta de mil documentos. Estes documentos foram extraídos a partir dos seguintes *sites* de negócios e notícias do mercado financeiro no ano de 2012: *Reuters*[1], *Bloomberg*[2], *Financial Times*[3], *Forbes*[4], *The New York Times*[5], *AllThingsD*[6] e *CNN Money*[7]. Foi definido um conjunto de entidades alvo para estudo (Apple, Google, Samsung, Microsoft, e Nokia) e selecionadas apenas páginas que citam, pelo menos, uma destas entidades.

A partir do conjunto de páginas contendo as entidades alvo, um subconjunto aleatório de 1.000 páginas foi selecionado e rotulado por 40 anotadores humanos. Cada anotador recebeu uma quantidade de 25 páginas, que foram avaliadas de acordo com sua polaridade

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

global e de acordo com a polaridade de cada entidade alvo presente na página. Do total de 1.000 documentos, 300 foram avaliados como positivo, 85 como negativo e 615 como neutro. A Tabela 4.1 apresenta a distribuição das polaridades por entidade, bem como o total de páginas associado a cada entidade.

Entidade	Positivo	Negativo	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
Totais	531	284	1301	2117

Tabela 4.1: Distribuição de polaridade por página.

Note nesta tabela que a distribuição é significativamente inclinada em ambas: polaridade (61% são neutros e apenas 13% são negativos) e entidade (Apple foi citada em 95% dos documentos, enquanto que Nokia em 12% apenas). A Tabela 4.2 apresenta a distribuição de entidades por página, onde cada entidade é uma companhia listada em *Forbes Fortune List 2012*[8]. Como podemos observar, a maioria dos documentos (617 que correspondem a 62%) citam mais de uma entidade. No entanto, o número mais comum de entidade por documento é de apenas um (38% dos documentos).

Entidades	Páginas
1	383
2	308
3	164
4	87
5	58

Tabela 4.2: Distribuição de entidade por página.

A Tabela 4.3 mostra a distribuição de polaridades por página. Embora a maioria dos documentos citem mais de uma entidade, em apenas 30% deles polaridades distintas são observadas.

[8]http://www.forbes.com/global2000/list/

Polaridades	Páginas
1	697
2	257
3	46

Tabela 4.3: Distribuição de polaridade por página.

4.2 Metodologia

Nossos experimentos foram realizados da seguinte forma. A coleção de dados rotulada foi processada de acordo com as estratégias de segmentação descritas na Seção 3.2. O texto foi segmentado em sentenças produzindo os conjuntos de dados correspondentes às estratégias SSM1, SSM2 e SSM3. Similarmente, CoreNLP foi usado para resolver anáfora e normalizar as citações relacionadas a cada entidade. Como esperado, resolução de anáfora foi a tarefa com custo de processamento mais alto. Os documentos foram então segmentados em sentenças para criar os conjuntos de dados correspondentes aos métodos SSM4, SSM5 e SSM6.

Depois de criar o conjunto de dados para cada entidade, foram realizadas as etapas adicionais de pré-processamento descritas por Azar [4]: (1) aplicação de *stemming* nas palavras; (2) remoção de *stopwords* usando a lista fornecida pela *WordNet*; (3) remoção de palavras que ocorrem menos que três vezes na coleção, uma vez que são consideradas palavras com pouca importância.

O desempenho de todos os modelos gerados foram avaliados por meio da métrica de acurácia (Seção 2.4) e todos os experimentos foram realizados utilizando validação cruzada de 5-*folds* (Seção 2.4). Assim, os resultados apresentados na seção a seguir referem-se a média da acurácia dos 5-*folds*.

4.3 Resultados

Nesta seção, apresentamos os resultados dos experimentos realizados. Inicialmente, expomos os critérios considerados na escolha do *baseline*, e então mostramos o impacto nos resultados ao aplicar nosso método.

4.3.1 Baseline

Adotamos como *baseline* o método supervisionado proposto por Azar [4], que também explora análise de polaridade no domínio financeiro. Primeiramente, avaliamos este método usando os documentos anotados com suas polaridades globais como conjunto de treino. Esta é uma configuração mais similar à usada por Azar, já que, na prática, ele considera que os documentos têm uma única entidade alvo e, consequentemente, uma única polaridade. Além disso, baseada na ideia apresentada em [31], separamos, para cada entidade E, apenas os documentos que citam E. Usando estes cinco conjuntos de documentos, construímos cinco modelos, um para cada entidade. Sendo assim, o modelo associado à entidade E usa como exemplos de treino as páginas que citam E. Isto corresponde ao método DbM, descrito na Seção 3.1.

Também fizemos comparação o método léxico, não supervisionado, proposto por Moilanen e Pulman [19] ao qual nos referimos neste texto como M&P09. Apesar dos autores não disponibilizarem a gramática de sentimentos necessária para a implementação deste método, nós obtivemos a licença da ferramenta (*TheySay API*[9]) que permite a análise de polaridade e realizamos testes na base que utilizamos no nosso trabalho. O método também foi avaliado de acordo com as polaridades das entidades. A Tabela 4.4 mostra os resultados obtidos para este experimento. Observe que avaliamos os métodos de acordo com as polaridades das entidade.

Entidades	Global (%)	M&P09 (%)	Ganho (%)	DbM (%)	Ganho (%)
Apple	55,87	47,52	-14,95	58,49	4,68
Google	54,17	57,48	6,11	61,42	13,38
Samsung	52,46	48,21	-8,10	61,11	16,48
Microsoft	54,42	54,28	-0,26	71,54	31,45
Nokia	45,72	45,90	0,39	57,60	25,98

Tabela 4.4: Método proposto por Azar treinado com polaridades global (Global), polaridades por entidade (estratégia DbM) e método léxico M&P09, avaliado por entidade.

Comparado ao método global, o M&P09 obteve desempenho muito abaixo do esperado. Isto se deve ao fato do método ser muito tendencioso para polaridades negativas e positivas. Como a maior parte dos documentos da base são considerados neutros, seu

[9]http://www.theysay.io/

desempenho é ruim. Em comunicação pessoal com os autores, fomos informados que, em sua configuração padrão, o M&P09 foca em estados gerais de coisas do mundo real que são positivas (por exemplo, ganhar um jogo de futebol ou ter um feriado) ou negativas (por exemplo, ficar doente ou hospitalizado), procurando sempre detectar mesmo os sinais mais fracos que indiquem tais polaridades. Baseado nesse primeiro resultado, podemos concluir que métodos puramente léxicos não são muito robustos para determinar a polaridade de entidades individuais.

Quanto ao método proposto por Azar observamos que, como esperado, usar múltiplos modelos, um por entidade, é melhor que usar um único modelo global. O menor ganho obtido foi para Apple. Em geral, quanto menor o número de documentos, maior é o ganho. Isto sugere que o aprendizado por entidade foi melhor pra aprender padrões de entidades menos populares, ao escapar dos vícios característicos das entidades mais populares. Por exemplo, da coleção de mil documentos, Apple é a mais citada, entretanto diversas citações se devem apenas ao fato dela ser a fabricante do Iphone ou Ipad. A partir dos resultados podemos concluir que um método baseado em múltiplos modelos, que considera as polaridades das entidades, é melhor que um método que infere polaridade globalmente.

Com base nos resultados obtidos, na próxima seção usaremos como *baseline* o método DbM.

4.3.2 Métodos Baseados em Segmentação por Sentença

Nesta seção, avaliamos nossos métodos que usam apenas os conjuntos de sentenças que citam as entidades de interesse como exemplos de treino. Em particular, comparamos todos os nossos métodos propostos na Seção 3.2 com DbM, onde os documentos não são segmentados.

Os resultados destes experimentos são apresentados na Tabela 4.5. Nesta tabela, as linhas representam o *baseline* DbM e os métodos baseados em sentenças usando casamento de *strings* (SSM1 a SSM3) e resolução de anáfora (SSM4 a SSM6). Os resultados representam a acurácia obtida pelos modelos na tarefa de classificar documentos em posi-

tivos, negativos e neutros para as entidades Apple, Google, Samsung, Microsoft e Nokia. Para cada um dos métodos, nós também apresentamos o ganho (ou perda) em relação ao *baseline*.

Método	Apple%	G%	Google%	G%	Samsung%	G%	Microsoft%	G%	Nokia%	G%
DbM	58,49	-	61,42	-	61,11	-	71,54	-	57,60	-
SSM1	57,13	-2,33	61,67	0,41	59,63	-2,42	69,76	-2,49	53,60	-6,94
SSM2	61,95	**5,92**	62,86	**2,34**	63,79	**4,39**	69,07	**-3,45**	64,00	**11,11**
SSM3	57,66	-1,43	62,62	1,95	59,95	-1,90	68,35	-4,46	56,80	-1,39
SSM4	53,47	-8,59	59,29	-3,47	61,65	0,88	67,64	-5,45	55.20	-4,17
SSM5	59,66	**2,01**	64,05	**4,28**	64,01	**4,74**	69,06	**-3,47**	60.00	**4,17**
SSM6	54,73	-6,43	60,95	-0,76	60,16	-1,56	67,64	-5,45	51.20	-11,11

Tabela 4.5: Método baseado em segmentação por sentença (SSM1, SSM2, SSM3, SSM4, SSM5 e SSM6) versus método baseado em documento (DbM).

Observamos na Tabela 4.5 que SSM2 (sentenças que não citam nenhuma entidade são descartadas) apresentou maiores ganhos em relação ao DbM para todas as entidades, exceto para Microsoft. Um comportamento similar observamos para o método correspondente baseado em resolução de anáfora (SSM5). Apesar do alto custo, os resultados para as estratégias que usam resolução de anáfora, em geral, não foram melhores. Os resultados para os outros métodos foram piores.

Em geral, a partir destes conjuntos de experimentos, podemos concluir que métodos que usam segmentação por sentença não são melhores que métodos baseados em documentos, mesmo quando usando técnicas complexas de PLN como resolução de anáfora. Os únicos métodos de segmentação que foram consistentemente melhores, aqueles que descartam sentenças que não citam entidades, apresentaram ganhos pequenos. Para todos os demais casos, os métodos de segmentação por documentos apresentaram resultados satisfatórios para a tarefa de classificação de polaridade com múltiplas entidades.

Como entre as estratégias de descarte, SSM2 e SSM5, o SSM2 tem custo menor, na próxima seção ele será usado como base de comparação.

4.3.3 Métodos Baseados em Detecção de Subjetividade

Nesta seção, aplicamos uma técnica de classificação hierárquica, em duas etapas. Na primeira etapa, o classificador separa os documentos neutros dos não-neutros. Ou seja, a

primeira etapa corresponde à tarefa de detecção de subjetividade. Na segunda etapa, os documentos classificados como não-neutros são separados em positivos ou negativos. Ou seja, a segunda etapa corresponde a uma classificação binária de polaridade.

Para a primeira etapa, utilizamos o mesmo método empregado em SSM2, treinado para classes neutra e não neutra. Para a segunda etapa, podemos usar dois diferentes classificadores: (a) o mesmo método empregado na primeira etapa, treinado com classes positivas e negativas, e (b) o método não supervisionado M&P09. Como vimos na Seção 4.3.1, o M&P09 não foi competitivo justamente por ser tendencioso para sentimentos não-neutros. Desta forma, ele se torna uma opção viável como classificador de segunda fase.

A Tabela 4.6 apresenta os resultados dos métodos apresentados nesta seção com seus respectivos ganhos em relação ao método SSM2. O método hierárquico, supervisionado nas duas fases, é chamado SSM2-H. O método hierárquico com segunda fase não supervisionada é chamado M&P09-H.

Entidade	SSM2 (%)	M&P09-H (%)	Ganho (%)	SSM2-H (%)	Ganho (%)
Apple	61,95	59,12	-4,57	61,11	-1,36
Google	62,86	62,14	-1,14	63,33	0,75
Samsung	63,79	59,05	-7,43	64,09	0,48
Microsoft	69,07	66,55	-3,65	66,90	-3,14
Nokia	64,00	61,60	-3,75	61,60	-3,75

Tabela 4.6: Método baseado em segmentação por sentença (SSM2) versus métodos baseados em detecção de subjetividade (M&P09-H e SSM2-H)

Como observado na Tabela 4.6, de forma geral, as estratégias hierárquicas não foram capazes de produzir melhores resultados que a não hierárquica. No caso do M&P09-H, contudo, o erro observado diminuiu significativamente em relação aos resultados originais, o que era esperado. Embora ainda inferior à estratégia supervisionada, note que qualquer esforço não supervisionado tem a grande vantagem de escalar melhor por não precisar de exemplos rotulados. Quanto ao SSM2-H, observamos resultados não satisfatórios. Ao analisar estes resultados em detalhe, verificamos que o método ganha na primeira fase da classificação, reconhecendo melhor os casos neutros. Apesar disso, ele perde significativamente na segunda pois, ainda que o classificador da primeira fase seja

melhor em detectar neutros, ele ainda comete muitos erros. Como resultado, muitos neutros vão para a segunda etapa. Isto se deve à distribuição de polaridade da base que é muito inclinada para os neutros, que representam quase 70% dos documentos.

Estes resultados nos levam a pensar que um método semi-hierárquico é mais adequado para este problema. Ou seja, ao invés de usar um classificador binário de polaridade na segunda fase, o correto seria usar novamente um classificador para as três classes. Esta é uma forma de reduzir o erro da segunda fase, uma vez que ainda são observados casos neutros. Avaliamos esse novo classificador, chamado SSM2-SH, na Tabela 4.7.

Entidade	DbM (%)	SSM2 (%)	SSM2-SH (%)	Ganho (%) sobre DbM	Ganho (%) sobre SSM2
Apple	58,49	61,95	62,89	7,52	1,52
Google	61,42	62,86	64,76	5,44	3,03
Samsung	61,11	63,79	65,58	7,31	2,80
Microsoft	71,54	69,07	70,11	-2,00	1,50
Nokia	57,60	64,00	65,60	13,89	2,50

Tabela 4.7: Método baseado em segmentação por sentença (SSM2) versus método baseado em detecção de subjetividade SSM2-SH. Resultados obtido com método DbM também são exibidos, para referência.

Nesta abordagem, os ganhos sobre o método sem segmentação, DbM, já são mais relevantes, especialmente, se considerarmos as entidades mais populares. Em relação ao SSM2, o SSM2-SH obteve ganhos melhores que o SSM2-H, mas ainda modestos. Apesar de construirmos um modelo na segunda fase que treina com exemplos para as três classes, este foi treinado com a coleção inteira. No entanto, após uma análise cuidadosa dos resultados nas duas etapas hierárquicas, verificamos que na segunda fase o número de documentos neutros é reduzido, em média, em 25% em relação à primeira fase. Isto implica que treinar o modelo com toda a coleção, que tem quase 70% dos documentos neutros, não reflete a distribuição dos dados da segunda fase, onde esse número de neutros é bem menor.

Uma forma de corrigir este problema seria usar a coleção de treino para estimar qual a distribuição final de positivos, negativos e neutros após a aplicação do classificador da primeira etapa. Isso poderia ser feito com uma validação cruzada dos casos de treino com o intuito de determinar que instâncias o classificador de primeira etapa consideraria não

neutras. Estas instâncias poderiam ser usadas como treino para o novo classificador de segunda etapa.

A partir do que foi apresentado, podemos concluir que, apesar dos resultados dos métodos usando abordagem hierárquica não terem obtido ganhos satisfatórios, esta é uma abordagem interessante para a tarefa de classificação de polaridade, principalmente se aplicada a cenários em que a coleção de dados apresenta uma distribuição mais balanceada. Além disso, acreditamos que para melhorar o desempenho deste tipo de abordagem o ideal é considerar sempre, no próximo nível de classificação, o erro gerado no nível anterior, garantindo assim a eficácia do próximo classificador.

Capítulo 5

Conclusão

Neste capítulo, resumimos as questões abordadas neste trabalho e apresentamos nossas conclusões finais. Expomos as limitações encontradas no decorrer do trabalho e sugerimos novas ideias de pesquisa que abordam questões deixadas em aberto por nosso estudo.

5.1 Resultados Obtidos

Neste trabalho, estudamos como melhorar o desempenho de um modelo de classificação na tarefa de classificar polaridades em documentos financeiros com múltiplas entidades. Para tanto, (a) propomos uma abordagem supervisionada baseada em múltiplos modelos, com o intuito de classificar a polaridade de textos com múltiplas entidades, (b) analisamos o impacto de estratégias de segmentação em texto, baseadas em heurísticas de casamento de *string* e resolução de anáfora, (c) estudamos a viabilidade do uso de abordagem léxica não supervisionada e por fim (d) propusemos um método de classificação hierárquica baseada em detecção de subjetividade.

Nossos resultados mostraram que uma abordagem baseada em múltiplos modelos é capaz de obter ganhos significativos sobre uma abordagem baseada em modelo global na tarefa de classificação de polaridade com múltiplas entidades. A segmentação do documento em sentenças que mencionam as entidades e a adoção de uma estratégia hierárquica também obtiveram ganhos, embora modestos.

Os experimentos que realizamos nos permitiram avaliar todas as questões de pesqui-

sas que motivaram este trabalho, apresentadas na Seção 1.2. Com base nos resultados, obtivemos as seguintes respostas:

- *Um método baseado em múltiplos modelos (um por entidade) é melhor que um método que infere polaridade globalmente?* Sim, é melhor. Observamos que a perda ao considerar um modelo global no aprendizado por entidade é alta devido ao número relativamente grande de documentos com múltiplas entidades. Modelos específicos são capazes de capturar peculiaridades destas entidades que se perdem quando consideramos modelos globais. Esse efeito é particularmente mais notável em entidades menos populares, já que o modelo global é enviesado para as entidades mais populares.

- *Qual o impacto das técnicas de segmentação de texto para este problema? Existe alguma técnica em particular, como heurísticas de casamento (matching) de strings e técnicas de resolução de anáfora, que melhore o desempenho do classificador?* A partir do conjunto de resultados que obtivemos, observamos que técnicas de segmentação por sentença, no geral, não melhoram o desempenho do classificador, exceto as estratégias que descartam sentenças que não citam nenhuma entidade (SSM2 e SSM5), onde obtivemos resultados consistentemente melhores. Técnicas que usam heurística de casamento de *strings* (SSM2) têm menor custo comparado com resolução de anáfora (SSM5), portanto são mais adequadas.

- *Um método que infere polaridade em múltiplos níveis (um para detectar opinião, seguido de outro para classificar polaridades positivas e negativas) é melhor que um método de uma única fase que classifica as três polaridades? Como este método deveria explorar a coleção de dados, em relação à distribuição das polaridades, de forma a maximizar o desempenho do classificador?* A princípio um método que infere polaridade em múltiplos níveis não é melhor que um método de uma única fase. Entretanto, é possível melhorar os resultados aplicando uma estratégia semi-hierárquica. Contudo, novos experimentos devem ser realizados considerando para a segunda fase, a distribuição final de positivos, negativos e neutros da coleção de

treino, após a aplicação do classificador da primeira fase.

- *Quão competitivo é um método não supervisionado quando comparado a um supervisionado?* Como esperado, mesmo um sofisticado método léxico, não supervisionado, não é tão eficaz quanto um método completamente supervisionado na tarefa de classificação de polaridade em documentos com múltiplas entidades. No entanto, estes se tornam mais competitivos quando combinados com métodos supervisionados em uma abordagem hierárquica. De um ponto de vista prático, eles serão sempre uma alternativa a considerar pelo fato de não dependerem de dados rotulados, que são extremamente difíceis de obter em muitos cenários.

5.2 Contribuições

Uma versão preliminar dos resultados apresentados neste trabalho foi publicado no XX Simpósio Brasileiro de Sistemas Multimídia e Web em Novembro de 2014 [12]. O artigo foi indicado para o melhor da conferência, tendo ganho o Prêmio de Menção Honrosa na trilha artigos completos.

5.3 Limitações

As principais limitações deste trabalho estão relacionadas à coleção de dados utilizada, uma vez que é relativamente pequena e muito desbalanceada. O fato da base ser constituída, em sua maior parte, por documentos neutros, dificulta o aprendizado de polaridades positivas e negativas. Nessas condições, é difícil avaliar diferentes estratégias, como as hierárquicas. Além disso, embora nosso objetivo fosse o foco em documentos financeiros, isto limita a universalidade das conclusões obtidas. Mesmo que tais conclusões sejam válidas em outros domínios, se faz necessário a avaliação de outras coleções.

No mais, inferimos polaridades para um conjunto pré-determinado de entidades. Em um cenário real, o número de entidades pode ser muito maior, tornando difícil a criação de modelos específicos. Tais entidades também devem ser determinadas de forma automática

ou serem representativas de uma seleção maior.

Como baseline léxico, não supervisionado, usamos um algoritmo estado-da-arte. Contudo não tivemos acesso aos algoritmos diretamente, mas sim através de uma API. Enquanto este é um software robusto, de uso comercial, ele não nos ofereceu flexibilidade suficiente para a sua melhor parametrização. Assim, os resultados obtidos foram, de certa forma, limitados ao conjunto de opções disponíveis.

5.4 Trabalhos Futuros

Neste trabalho, assumimos que as polaridades das diferentes entidades são independentes, o que provavelmente não ocorre. Por exemplo, em nossa coleção, quando um documento é positivo para Apple é mais provável que seja negativo para Samsung do que um documento aleatório. Da mesma forma, entidades que colaboram entre si tendem a ter mais polaridades em comum que diferentes (como nos casos 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, estas informações podem ser usadas no processo de aprendizado.

Assim, no futuro pretendemos investigar técnicas que considerem a influência mútua das polaridades. Em particular, estudaremos técnicas discutidas em [9], tal como *stacking* e regressão multi-variada com saídas correlacionadas em lugar de uma classificação. Neste último caso, o alvo da previsão seria um valor numérico entre -1 e +1 (onde zero indica uma polaridade neutra) para cada entidade.

Outro trabalho futuro é a investigação mais detalhada de novos métodos de classificação semi-hierárquica, onde a definição das amostras de treino para as etapas mais específicas considere a taxa de erro das etapas mais gerais.

Em relação às limitações apresentadas, pretendemos verificar nossas conclusões em coleções de outros domínios e estudar métodos híbridos baseados em múltiplos 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 se-

gunda 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).

Referências Bibliográficas

[1] ABDUL-MAGEED, M., DIAB, M. T., AND KORAYEM, M. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (2011), Association for Computational Linguistics, pp. 587–591.

[2] ABDUL-MAGEED, M., KÜBLER, S., AND DIAB, M. Samar: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (2012), Association for Computational Linguistics, pp. 19–28.

[3] ARAÚJO, M., GONÇALVES, P., AND BENEVENUTO, F. Measuring sentiments in online social networks. In *Proceedings of the 19th WebMedia* (2013), ACM Press, pp. 97–104.

[4] AZAR, P. *Sentiment Analysis Financial News*. PhD thesis, Hardvard College, 2009.

[5] BAEZA-YATES, R. A., AND RIBEIRO-NETO, B. A. *Modern Information Retrieval, Second Edition*. Pearson Education Ltd., Harlow, England, 2011.

[6] BECKER, K., AND TUMITAN, D. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. *Simpósio Brasileiro de Banco de Dados* (2013).

[7] BOLLEN, J., MAO, H., AND ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (2011), 1–8.

[8] CHANG, C.-C., AND LIN, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. 2*, 3 (May 2011), 27:1–27:27.

[9] DEMBCZYŃSKI, K., WAEGEMAN, W., CHENG, W., AND HÜLLERMEIER, E. On label dependence in multi-label classification. In *Workshop Proceedings of Learning from Multi-Label Data* (Haifa, Israel, 2010), pp. 5–12.

[10] DEVITT, A., AND AHMAD, K. A lexicon for polarity: Affective content in financial news text. *Proceedings of Language For Special Purposes* (2007).

[11] DEVITT, A., AND AHMAD, K. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (2007).

[12] FERREIRA, J. Z., RODRIGUES, J., CRISTO, M., AND DE OLIVEIRA, D. F. Multi-entity polarity analysis in financial documents. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web* (2014), ACM, pp. 115–122.

[13] GRYC, W., AND MOILANEN, K. Leveraging textual sentiment analysis with social network modelling. In *Proc. of the "From Text to Political Positions"Workshop (T2PP)* (2010).

[14] IM, T. L., SAN, P. W., ON, C. K., ALFRED, R., AND ANTHONY, P. Analysing market sentiment in financial news using lexical approach. *2013 IEEE Conference on Open Systems (ICOS)* (2013), 145–149.

[15] LIU, B. Sentiment analysis and subjectivity. *Handbook of natural language processing 2* (2010), 627–666.

[16] LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies 5*, 1 (2012), 1–167.

[17] MANNING, C. D., AND SCHÜTZE, H. *Foundations of statistical natural language processing*. MIT press, 1999.

[18] MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J., BETHARD, S. J., AND MCCLOSKY, D. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2014), pp. 55–60.

[19] MOILANEN, K., AND PULMAN, S. Multi-entity sentiment scoring. In *RANLP* (2009), G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, Eds., RANLP 2009 Organising Committee / ACL, pp. 258–263.

[20] MONTEJO-RÁEZ, A., MARTÍNEZ-CÁMARA, E., MARTÍN-VALDIVIA, M. T., AND UREÑA-LÓPEZ, L. A. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language 28*, 1 (2014), 93–107.

[21] MOURAD, A., AND DARWISH, K. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2013), pp. 55–64.

[22] NG, V. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2010), ACL '10, pp. 1396–1411.

[23] PANG, B., AND LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (2004), Association for Computational Linguistics, p. 271.

[24] PANG, B., AND LEE, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval 2*, 1-2 (2008), 1–135.

[25] PANG, B., AND LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr. 2*, 1-2 (Jan. 2008), 1–135.

[26] PANG, B., LEE, L., AND VAITHYANATHAN, S. 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).

[27] POESIO, M., PONZETTO, S., AND VERSLEY, Y. Computational models of anaphora resolution: A survey. *Linguistic Issues in Language Technology* (2011).

[28] REFAEE, E., AND RIESER, V. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme* (2014), p. 16.

[29] ROMANYSHYN, M. Rule-based sentiment analysis of ukrainian reviews. *International Journal of Artificial Intelligence & Applications 4*, 4 (2013).

[30] SCHUMAKER, R. P., AND CHEN, H. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst. 27* (March 2009), 12:1–12:19.

[31] WARD, C. B., CHOI, Y., SKIENA, S., AND XAVIER, E. C. Empath: A framework for evaluating entity-level sentiment analysis. In *Emerging Technologies for a Smarter World (CEWIT)* (2011), IEEE, pp. 1–6.

[32] WILSON, T., WIEBE, J., AND HOFFMANN, P. 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).

[33] WITTEN, I. H., FRANK, E., AND HALL, M. A. *Data mining : practical machine learning tools and techniques*, 3rd ed. Morgan Kaufmann, San Francisco, CA, USA, 2011.

[34] YI, J., NASUKAWA, T., BUNESCU, R., AND NIBLACK, W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on* (November 2003), pp. 427 – 434.

[35] YU, H., AND HATZIVASSILOGLOU, V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (2003), Association for Computational Linguistics, pp. 129–136.