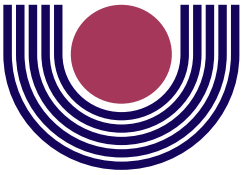

Unioeste - Universidade Estadual do Oeste do Paraná
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
Colegiado de Ciência da Computação
Curso de Bacharelado em Ciência da Computação

Mineração de Opinião Baseada em Extração de Aspectos

Thales Felipe Costa Bertaglia

CASCAVEL
2015

Thales Felipe Costa Bertaglia

Mineração de Opinião Baseada em Extração de Aspectos

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel

Orientador: Prof. Dr. Clodis Boscarioli

CASCAVEL
2015

Thales Felipe Costa Bertaglia

MINERAÇÃO DE OPINIÃO BASEADA EM EXTRAÇÃO DE ASPECTOS

Monografia apresentada como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. Dr. Clodis Boscarioli
Colegiado de Ciência da Computação,
UNIOESTE

Prof. Dr. Marcio Seiji Oyamada
Colegiado de Ciência da Computação,
UNIOESTE

Prof. Dra. Sarajane Marques Peres
Escola de Artes, Ciências e Humanidades, USP

Cascavel, 12 de fevereiro de 2016

DEDICATÓRIA

Ao meu avô Enelvo Bertaglia, in memoriam.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API	*Application Programming Interface*
PLN	Processamento de Linguagem Natural
NER	*Named-Entity Recognition*
MD	Mineração de Dados
SVM	*Support Vector Machine*
PMI	*Pointwise Mutual Information*
HMM	*Hidden Markov Model*
pLSA	*Probabilistic Latent Semantic Analysis*
LDA	*Latent Dirichlet Allocation*
NLTK	*Natural Language Toolkit*
XML	*Extensible Markup Language*
POS	*Part-of-speech*

Sumário

Resumo

Análise de sentimentos, ou mineração de opinião, é o estudo computacional de opiniões, sentimentos, avaliações e atitudes em relação a entidades expressadas em documentos textuais. O principal objetivo da análise de sentimentos é extrair opiniões e sentimentos relacionados a elas de fontes como avaliações de consumidores, *blogs* e fóruns de discussão. Neste trabalho são apresentados e discutidos os principais conceitos, técnicas e abordagens computacionais da área. Também é especificado um sistema com o objetivo de extrair opiniões sobre características de produtos dadas em avaliações de consumidores. O sistema é composto pelos módulos de pré-processamento, extração de aspectos, extração de sentimentos e sumarização. Cada um dos módulos é discutido, assim como os algoritmos implementados. O algoritmo de extração de aspectos é avaliado experimentalmente visando mensurar sua precisão. Cada etapa que o constitui é avaliada separadamente de modo a verificar seu impacto na saída final do algoritmo Por fim, o sistema é analisado como um todo e os resultados obtidos são comparados aos já consolidados na literatura.

Palavras-chave: Análise de Sentimentos, Mineração de Opinião, Extração de Aspectos.

Capítulo 1

Introdução

As opiniões e as experiências de outras pessoas constituem uma importante fonte de informação em nossa vida. É comum buscar-se recomendações de conhecidos sobre qual celular ou computador comprar, qual restaurante ir e até mesmo qual médico consultar. Atualmente, avaliações de consumidores (*reviews*) constituem-se em um recurso valioso para auxiliar na tomada de decisão [Bross 2013]. Além de ajudar o consumidor a decidir na hora de efetuar compras, as *reviews* oferecem *feedback* gratuito e espontâneo aos fabricantes e às empresas, visto que os clientes escrevem as avaliações sem obrigação e podem exprimir opiniões sem restrição.

No entanto, o grande número de avaliações disponíveis *online* dificulta o processamento e a compreensão total das informações. Podem existir milhares de *reviews* escritas sobre produtos populares – uma quantia não muito viável para humanos lerem. Desse modo, há uma tendência em desenvolver sistemas que possam automaticamente extrair informações em avaliações de consumidores. Essa tarefa é chamada de análise de sentimentos ou mineração de opinião.

A análise de sentimentos não se restringe ao domínio de avaliações de consumidores. Ela pode ser aplicada em vários cenários – como debates políticos, sugestões de produtos e análise de tendências de mercado. Qualquer contexto que envolva opiniões sendo expressadas pode ser uma aplicação. A análise de sentimentos surge como uma alternativa para lidar com a grande quantia de opiniões disponíveis, visando facilitar a compreensão das informações contidas nelas. Seu principal objetivo é identificar automaticamente opiniões expressadas em textos e os sentimentos relacionados a elas. Sentimento, nesse contexto, é o que o emissor da opinião pensa sobre a entidade avaliada. Um usuário que escreve uma avaliação dizendo que "*a qualidade desse celular é péssima*", por exemplo, demonstra um sentimento negativo em relação ao produto.

A identificação de opiniões pode ser feita de diversas formas e em diferentes níveis de análise. Pode-se considerar que a avaliação toda contém uma opinião, que cada frase do texto expressa um sentimento ou que cada característica avaliada tem uma opinião relacionada. Do ponto de vista computacional, existem diferentes abordagens para a solução do problema, sendo que elas podem ser basicamente divididas em técnicas de aprendizado supervisionado e não-supervisionado.

O objetivo desse trabalho é apresentar os principais conceitos de análise de sentimentos e algumas técnicas utilizadas para efetuá-la. O trabalho é orientado a opiniões expressas em avaliações de consumidores disponíveis em *websites* de lojas (como a Amazon), com a implementação de um sistema para aplicar conceitos de análise de sentimentos na prática.

Com base nesses conceitos e técnicas, propõe-se um sistema de extração de opinião em nível de entidade e aspecto, que será discutido no texto. Esse sistema tem como objetivo identificar automaticamente opiniões em avaliações de consumidores sobre produtos. As opiniões identificadas não se restringem à avaliação geral – cada característica do produto avaliada é considerada importante e deve ser identificada. O processo de identificação dessas característica é denominado extração de aspectos. O algoritmo para extração de aspectos implementado no sistema é avaliado experimentalmente visando mensurar sua precisão. Cada etapa que o constitui é avaliada separadamente de modo a verificar seu impacto na saída final do algoritmo. O sentimento em relação a cada uma dessas características também é identificado. Por fim, são gerados resumos de opinião que visam comprimir a informação para facilitar a compreensão e reduzir o tempo necessário de leitura. Todas as avaliações utilizadas são escritas em língua inglesa, uma vez que a implementação usa técnicas de processamento de linguagem natural específicas para esse idioma.

Esse documento está organizado da seguinte forma:

- **Capítulo 2**: apresenta de forma conceitual o campo de pesquisa da análise de sentimentos. Os principais conceitos são discutidos.

- **Capítulo 3**: apresenta mais detalhadamente o processo de mineração de opinião, explicando os diferentes níveis de análise. Nesse capítulo também são apresentadas brevemente as principais técnicas computacionais utilizadas na área.

- **Capítulo 4**: apresenta o sistema proposto, detalhando cada um dos módulos que o compõe e os algoritmos implementados.
- **Capítulo 5**: apresenta a metodologia e as bases de dados utilizadas para testes, e discute os resultados obtidos a partir do sistema.
- **Capítulo 6**: relata as conclusão do projeto e apresenta perspectivas de novos trabalhos.

Capítulo 2

Mineração de Opinião e Análise de Sentimentos

Análise de sentimentos, também chamada de mineração de opinião, é o campo de pesquisa que estuda opiniões, sentimentos, avaliações, atitudes e emoções de pessoas em relação a determinadas entidades – como produtos, serviços, organizações, eventos e assuntos, bem como seus atributos constituintes [Liu 2012]. É uma área ampla, pois pode ser aplicada a qualquer domínio que contenha opiniões, e abrange vários problemas – como extração de informação, classificação e sumarização textual. Existem diferentes terminologias encontradas na literatura, como *mineração de opinião*, *análise de sentimentos*, *extração de opinião*, *mineração de sentimentos*, *análise de subjetividade*, entre outros. No entanto, todos esses termos estão contidos na área de mineração de opinião e análise de sentimentos. Segundo [Pang e Lee 2008], a diversidade de termos para denominar a área surgiu devido à diferença de *background* dos pesquisadores – que se originam de diferentes áreas. Nesse trabalho, os termos *mineração de opinião* e *análise de sentimentos* são usados sem distinção.

Há muitas razões pelas quais a mineração de opinião é uma tarefa a ser explorada na área da Ciência da Computação. Nas seções seguintes serão discutidos alguns cenários de grande valia nos quais técnicas de mineração de opinião podem ser eficientemente empregadas.

2.1 Motivação

Opiniões são pontos-chave no processo de decisão e no comportamento das pessoas. É comum que consumidores busquem a opinião de outros indivíduos sobre um produto antes de adquiri-lo. Também é comum buscar recomendações sobre serviços ou empresas – seja sobre

qual loja comprar um produto ou até mesmo em qual médico se consultar. Para empresas, é importante saber a opinião pública sobre seus produtos ou serviços prestados. O uso de mineração de opinião será discutido no contexto dos consumidores e no das organizações.

2.1.1 Mineração de Opinião para Consumidores

Um consumidor, nesse contexto, é qualquer usuário final que deseja analisar opiniões sobre uma entidade para auxiliar em seu processo decisório a respeito dessa entidade. Pessoas lendo avaliações sobre um produto em uma loja *online*, buscando informações sobre um candidato político em um fórum ou lendo relatos de viagem em uma rede social são todas consumidores de opinião.

Segundo um estudo conduzido pelo The Kelsey Group [Group 2015], uma empresa de análise estratégica focada em propaganda, consumidores americanos estão dispostos a pagar até 20% a mais por produtos avaliados como "Excelentes"ou com cinco estrelas em avaliações *online*. O estudo examina o impacto que as avaliações *online* de consumidores causa nas vendas e é baseado em questionários aplicados a mais de dois mil americanos usuários da Internet. Foram consideradas avaliações a respeito de restaurantes, hotéis, viagens e serviços automotivos, legais, médicos e domésticos. A Tabela 2.1, adaptada de [Group 2015], indica a porcentagem de usuários que identificou que as avaliações tiveram impacto significativo em sua compra ou contratação de serviço.

Tabela 2.1: Porcentagem de usuários que identificou que as avaliações tiveram impacto significativo em sua compra

Serviço	Porcentagem de Usuários
Restaurantes	79%
Hotéis	87%
Viagens	84%
Automotivos	78%
Domésticos	73%
Médicos	76%
Legais	79%

Os dados obtidos ressaltam a importância das avaliações no processo de compra dos consumidores. No entanto, outro estudo sobre compras *online* [Horrigan 2015] indica que consumidores nem sempre consideram as informações disponíveis suficientemente claras. Segundo o

estudo, mais de metade dos usuários tem de lidar com situações frustantes ao realizar compras *online*.

- 43% dos usuários já se sentiram **frustrados** pela falta de informação ao buscar ou comprar produtos *online*.
- 32% se sentiram **confusos** pelas informações encontradas durante sua compra ou busca.
- 30% se sentiram **sobrecarregados** pela quantidade de informação encontrada.

Com a popularização da Internet, os consumidores têm a seu dispor uma quantia imensa de opiniões que são compartilhadas por meio de fóruns de discussão, *blogs*, vídeos e diversas redes sociais. Apesar de o compartilhamento de opinião ajudar no processo de decisão, muitas vezes o consumidor se depara com um número elevado de avaliações e não é capaz de analisar todas – fato que os deixa sobrecarregados em vez de auxiliar. A Figura 2.1 mostra um exemplo de avaliações feitas sobre um popular aparelho de celular, extraídas da loja virtual Amazon [Amazon 2015].

Figura 2.1: Exemplo de um número elevado de avaliações feitas sobre um produto

A figura indica um total de 1162 avaliações submetidas, sendo que cada uma pode ser constituída por várias linhas de texto. Essa quantia de informação é elevada e um consumidor gastaria muito tempo para ler todo o conteúdo disponível. Desse modo, o consumidor geralmente lê apenas algumas avaliações, porém pode assim perder informações que seriam relevantes a sua compra. Nesse contexto, a mineração de opinião provê técnicas que visam facilitar a aquisição de informações advindas de avaliações de produtos. Essas técnicas têm como objetivo extrair automaticamente as principais características expostas nas avaliações e as opiniões sobre elas.

Por fim, são gerados resumos que comprimem as informações obtidas de modo a gerar um texto breve e de compreensão rápida para os usuários – esse processo é chamado de sumarização textual [Wang e Ren 2015]. A mineração de avaliações de consumidores é o principal objetivo desse trabalho e as técnicas empregadas para tal fim serão detalhadas ao decorrer do texto.

A mineração de opinião não se restringe à análise de avaliações de consumidor. Qualquer fonte de opinião é um possível domínio de aplicação. Há exemplos de trabalhos que identificam opiniões baseando-se em *tweets* [Jiang et al. 2011], que recomendam automaticamente produtos baseando-se em publicações de usuários em redes sociais [Zoghbi, Vulić e Moens 2013] e que identificam preferências políticas [Bakliwal et al. 2013]. A diversidade de aplicações demonstra o potencial da área.

2.1.2 Mineração de Opinião para Organizações

Uma organização é considerada como um tipo de usuário que utiliza informação proveniente de opiniões como meio para atingir determinado objetivo. Uma empresa que analisa avaliações feitas por clientes a respeito de seus produtos e utiliza essas informações para melhorá-los é um tipo de organização. Adquirir opinião do público já é uma prática comum no meio empresarial. Organizações frequentemente conduzem pesquisas de opinião pública com fins de *marketing*, relações públicas e campanhas políticas. Essas pesquisas são conduzidas e aplicadas por instituições especializadas e usualmente têm custo elevado [PHD 2015]. Entretanto, com a quantia de opinião publicamente disponível, a mineração de opinião pode ser uma alternativa viável.

A mineração de opinião pode ser aplicada em vários tipos de sistemas inteligentes, porém é mais adequada quando utilizada em aplicações de *business intelligence* [Pang e Lee 2008]. Os sistemas de *business intelligence* utilizam os dados disponíveis nas organizações para disponibilizar informação relevante para a tomada de decisão e estão tradicionalmente associados às tecnologias de *data warehouse*, *on-line analytical processing* e *data mining* [Santos e Ramos 2006]. A análise de sentimentos pode ser inserida nesse tipo de sistema de maneira análoga a *data mining*.

Para exemplificar, o cenário descrito em [Pang e Lee 2008] é utilizado: uma grande fabricante de computadores está passando por uma crise devido à baixa quantia de vendas de seu novo modelo de computador. Decepcionada com a situação, a empresa se pergunta por que os

consumidores não compram seu computador. Ao passo que dados concretos sobre o produto (como preço, peso e modelos dos concorrentes) são relevantes, é necessário focar-se no ponto de vista pessoal dos consumidores para responder a essa questão. Características intangíveis como " o *design* é antiquado" ou "o serviço de atendimento ao consumidor não é amigável" são importantes para os consumidores e precisam ser consideradas. Para esse fim, é necessário lidar com a extração de opinião de documentos não-estruturados escritos por humanos – tarefa na qual tecnologias de análise de sentimentos são muito bem adequadas. Nesse contexto, é possível implementar um sistema de *business intelligence* com mineração de opinião que:

a) Encontre avaliações sobre o produto na Internet.

b) Crie versões condensadas de avaliações individuais ou um resumo dos principais pontos abordados nas avaliações.

c) Exiba as informações encontradas de maneira clara e eficiente.

O resultado desse processo pode, por fim, ser analisado por um especialista no domínio que irá então extrair o conhecimento que poderá ser aplicado na situação da empresa. Nesse caso, a mineração de opinião poupou o especialista de ter que ler potencialmente centenas de avaliações de consumidores relatando opiniões muitas vezes semelhantes. Além do cenário apresentado, a análise de sentimentos pode ser usada em *marketing* inteligente, sistemas de recomendação e anúncios personalizados, análise de tendências de mercado, detecção de *spam*, entre outros.

Aplicações industriais de sistemas de mineração de opinião têm surgido recentemente. Algumas corporações, como Google, Hewlett-Packard e Microsoft, também têm implementado soluções de análise de sentimentos [Liu 2012]. Algumas aplicações disponíveis *online* são as seguintes:

- **Twitrratr**: é uma ferramenta de busca para encontrar opiniões expressas na rede social Twitter, disponível em [Twtbase 2015]. O sistema retorna *tweets* divididos em categorias contendo opiniões positivas, negativas e neutras a respeito do termo de pesquisa informado. A Figura 2.2 ilustra a *interface* da ferramenta.

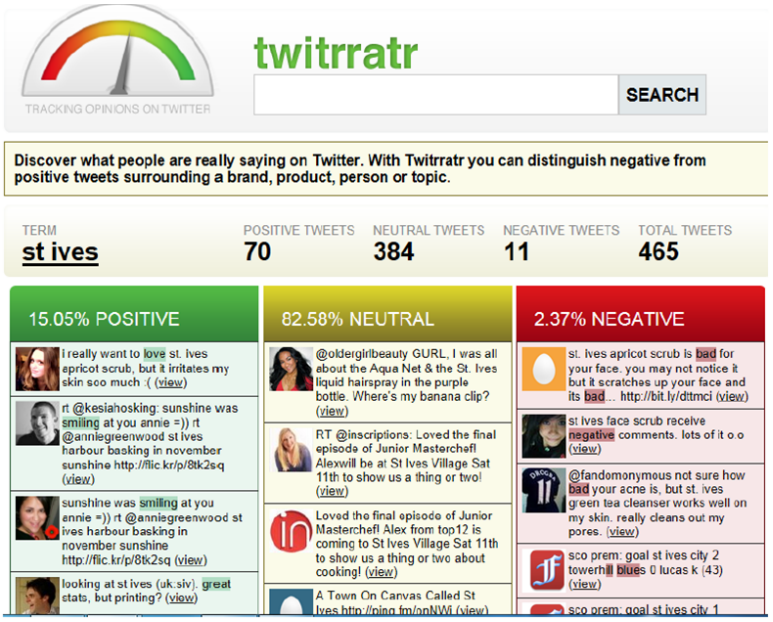

Figura 2.2: Ferramenta Twitrratr

- **Tweetfeel**: ferramenta, disponível em [Tweetfeel 2014], semelhante à Twitrattr mas que não filtra os *tweets* pela opinião – apenas os retorna e informa a polaridade de cada um. A Figura 2.3 ilustra sua *interface*.

Figura 2.3: Ferramenta Tweetfeel

- **Sentiment140**: ferramenta desenvolvida por pesquisadores da Universidade de Stanford, disponível em [Sentiment140 2015]. A Figura 2.4 ilustra sua *interface*. A ferramenta classifica *tweets* de acordo com a opinião expressa e faz análises estatísticas básicas sobre a distribuição da polaridade. Possui uma *Application Programming Interface* (API) que faz com que a ferramenta possa ser integrada a outros sistemas.

Figura 2.4: Ferramenta Sentiment140

Mesmo não possuindo funcionalidades complexas, essas ferramentas demonstram a utilidade e o potencial da mineração de opinião. Uma rede social como o Twitter, na qual usuários do mundo inteiro publicam em torno de 500 milhões de *tweets* por dia [Stats 2015], é uma fonte diversificada de opiniões sobre qualquer assunto e pode ser explorada por empresas para aplicações comerciais.

2.1.3 Mineração de Opinião para Aplicações de PLN

A área de análise de sentimentos está intimamente ligada ao processamento de linguagem natural. Pode-se considerar que a mineração de opinião é um sub-tópico de pesquisa advindo do PLN, visto que muitos de seus problemas são também questões de PLN. As duas áreas começaram a se desenvolver amplamente a partir do ano 2000. Segundo [Liu 2012], isso é em partes devido ao fato de que anteriormente havia poucos documentos de opinião disponíveis digitalmente. A partir de então, a análise de sentimento se tornou uma das áreas de pesquisa mais ativas em PLN. Desse modo, a mineração de opinião pode auxiliar no crescimento do PLN, principalmente em tarefas de:

1. **Tradução Automática de Texto**: A classificação de sentimentos *cross-language* é uma subárea da mineração de opinião e tem como objetivo efetuar a classificação de documentos de opinião em múltiplas línguas. Há dois motivos principais para realizar esse tipo de classificação [Liu 2012]. Primeiramente, pesquisadores de diferentes países desejam desenvolver sistemas de análise de sentimentos em suas próprias línguas. No en-

tanto, a pesquisa na área é centrada na língua inglesa – portanto há poucos recursos e ferramentas disponíveis em outras línguas. O segundo motivo é que em muitas aplicações, empresas gostariam de analisar e comparar opiniões sobre seus produtos e serviços em diferentes países. Para possibilitar esse tipo de análise, é necessário lidar com um fluxo contínuo de uma grande quantia de informação escrita em línguas distintas [Hogenboom et al. 2013]. A solução é aperfeiçoar e ampliar as técnicas de tradução automática de texto para que o uso de ferramentas disponíveis em inglês se torne viável também em outras línguas. Dessa maneira, a mineração de opinião contribuiu para o PLN. O trabalho de [Brooke, Tofiloski e Taboada 2009] propõe técnicas para adaptar uma calculadora de orientação semântica da língua inglesa para língua espanhola. Os resultados indicam que, apesar de a automação da tradução ser custosa, é possíveis obter resultados satisfatórios.

2. **Sumarização**: O processo de sumarização consiste na produção de uma versão compactada de um documento, composta pelas sentenças mais relevantes ao seu conteúdo [Jurafsky e Martin 2009]. Essa tarefa está diretamente relacionada à mineração de opinião, visto que parte de seu objetivo é gerar resumos que condensem informação obtidas por meio de documentos de opinião. Assim, sistemas de análise de sentimentos podem incentivar a área de PLN desenvolvendo técnicas de sumarização orientadas a opiniões.

3. **Indexação de Informação**: Sistemas de indexação de informação muitas vezes retornam grandes quantidades de documentos como resultado de uma busca feita por um usuário. Estes documentos nem sempre são todos relevantes para o usuário, o que provoca insatisfação [Paetzold 2013]. Considerando buscas realizadas sobre documentos de opinião, a análise de sentimentos pode prover respostas mais adequadas e que atendam de forma mais fiel aos parâmetros requisitados pelo usuário.

2.2 Conceitos de Análise de Sentimentos

Opiniões, diferente de informações factuais, têm uma característica importante – são subjetivas. Além disso, são escritas em linguagem natural de forma não-estruturada. É necessário, portanto, abstrair conceitos do problema de análise de sentimentos e defini-los. Essas defini-

ções são vitais para estruturar a pesquisa e reduzir a complexidade de se lidar com textos em linguagem natural.

Ao decorrer do texto, serão utilizadas principalmente opiniões oriundas de avaliações de consumidores sobre produtos para ilustrar e definir conceitos. Apesar de as ideias poderem ser aplicadas em outras fontes de opinião, avaliações sobre produtos apresentam opiniões de forma direta e focada, facilitando a visualização dos conceitos envolvidos.

2.2.1 Definição de Opinião

A seguinte avaliação de consumidor, retirada de [Walmart 2015], será utilizada para introduzir o problema:

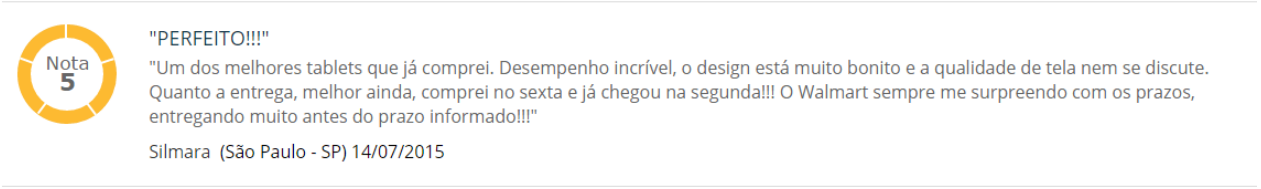

Figura 2.5: Exemplo de uma avaliação de consumidor sobre um *tablet*

Com base no texto da avaliação, é possível notar algumas características:

1. A avaliação possui várias opiniões, tanto sobre o produto em si (um *tablet*), quanto sobre a loja responsável pela venda. A frase "*desempenho incrível, o design está muito bonito e a qualidade de tela nem se discute*" apresenta opinião positiva sobre três atributos distintos do produto. A última frase, que trata sobre a loja, indiretamente diz que o serviço de entrega é rápido. Nota-se que uma opinião é constituída de dois componentes-chave: um alvo g e um sentimento s, ou seja, um par (g, s) no qual g pode ser qualquer entidade ou atributo de entidade e s o sentimento relacionado a g. Em geral, sentimentos são considerados como positivos, negativos ou neutros ou como algum valor numérico (por exemplo número de estrelas). Na Figura 2.5, uma nota de 0 a 5 é atribuída como avaliação da entidade em geral.

2. A avaliação apresenta a opinião de uma pessoa – o usuário Silmara. O emissor da opinião é conhecido na literatura como proprietário ou fonte da opinião, advindo dos termos em inglês *opinion holder* ou *opinion source* [Kim e Hovy 2004].

3. A data da avaliação é 14/07/2015. Essa informação é importante para avaliar o aspecto temporal das opiniões – como elas mudam com o passar do tempo, por exemplo.

Partindo disso, pode-se definir uma **opinião** como uma quádrupla (g, s, h, t), sendo g o alvo da opinião, s o sentimento em relação ao alvo, h a fonte da opinião e t o tempo (ou data) em que a opinião foi expressada [Liu 2012]. Mais componentes podem ser adicionados à tupla conforme a necessidade da aplicação, como idade e nacionalidade da fonte de opinião.

Aplicando essa definição à avaliação da Figura 2.5, tem-se que g é o produto, s é positivo, h é Silmara e t é 14/07/2015. No entanto, baseando-se nesse exemplo já é possível encontrar problemas na definição proposta. A questão principal é que o elemento g considera que a avaliação trata sobre um único alvo. Apesar de avaliar um produto específico, ela faz referência a diferentes características e até mesmo entidades. Desse modo, ao definir g é necessário generalizar todas as informações apresentadas e transformá-las em um alvo único. Essa generalização acarreta em perda de informação e portanto não reflete a verdadeira opinião expressada. Consequentemente, o sentimento s também é inválido, pois considera apenas o alvo genérico. Para resolver esse conflito, é necessário expandir o conceito de entidade-alvo.

Uma **entidade** e é um produto, serviço, tópico, assunto, pessoa, organização ou evento. Ela é representada como um par $e : (T, W)$, sendo T um conjunto hierárquico de partes da entidade e W um conjunto de atributos (ou características) de e [Liu, Mobasher e Nasraoui 2011]. Um aparelho celular, por exemplo, é uma entidade. Ele possui um conjunto de atributos – como *duração da bateria*, *qualidade da ligação* e *peso* – e um conjunto de partes – como *bateria*, *visor* e *touch screen*. *Bateria*, por sua vez, também possui um conjunto de atributos, assim como todos os elementos do grupo de partes. Essa representação hierárquica é frequentemente complexa demais para aplicações. Em problemas práticos, pode ser inviável identificar cada uma das partes da entidade e dos atributos de cada uma. Desse modo, com o propósito de simplificar a representação apenas dois níveis hierárquicos são considerados e o termo **aspecto** é utilizado para definir tanto atributos quanto partes constituintes.

Dada a definição de entidade, pode-se a partir dela definir **opinião** como uma quíntupla $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, sendo e_i o nome de uma entidade, a_{ij} um aspecto de e_i, s_{ijkl} o sentimento a respeito do aspecto a_{ij}, h_k a fonte da opinião e t_l o tempo em que a opinião foi expressada por h_k. Analogamente à definição anterior, s_{ijkl} pode ser positivo, negativo, neutro ou assumir

valor numérico.

Nessa definição, o índice subscrito tem como objetivo deixar claro que os componentes da quíntupla estão inter-relacionados e devem ser correspondentes. Além disso, também expressa o conceito de que uma entidade pode ter vários aspectos a_{ij}, pois o índice i indica a entidade e o índice j permite uma quantia qualquer de aspectos. Em geral, os cinco componentes são essenciais, não obter algum deles ocasiona perda de informações. No entanto, em casos específicos, dependendo da aplicação, é possível ignorar componentes que não forem considerados relevantes – por exemplo t_l em uma aplicação que não considera o aspecto temporal das opiniões. É importante ressaltar que essa definição é válida para opiniões regulares, aquelas em que a opinião sobre a entidade é dada diretamente. Existem também opiniões comparativas, aquelas em que a opinião é dada por meio da comparação entre entidades. Esse tipo de opinião deve ser tratado de maneira específica e não é considerado nesse trabalho. O trabalho de [Jindal e Liu 2006] oferece uma visão geral sobre o processo de mineração de opiniões comparativas.

Aplicando a nova definição de opinião ao exemplo da Figura 2.5, uma possível abstração é: e_1 = *tablet*, e_2 = loja. a_{11} = desempenho, a_{12} = *design*, a_{13} = qualidade da tela, a_{21} = prazo de entrega. h_1 = usuário Silmara. t_1 = 14/07/2015. s_{1111} = **positivo** ("*incrível*"), s_{1211} = **positivo** ("*muito bonito*"), s_{1311} = **positivo** ("*nem se discute*", sentimento implícito), s_{2111} = **positivo** ("*entregando muito antes do prazo informado*", sentimento implícito). Como só há opinião de uma fonte em uma única data, é possível omitir os dois últimos índices na definição de s.

2.2.2 Tipos de Opinião

Opiniões podem ser classificadas em relação a como tratam sobre a entidade alvo e a como são expressadas no texto. No primeiro caso, são dividas em *regulares* ou *comparativas*. No segundo, em *explícitas* ou *implícitas*.

- **Opiniões regulares e comparativas:** uma opinião regular, geralmente chamada apenas de opinião, trata somente da entidade que está sendo avaliada e pode ser dividida em *direta* ou em *indireta*. A opinião é dita *direta* quando é expressada diretamente sobre a entidade ou um aspecto da entidade – "*a duração da bateria é ótima*", por exemplo, avalia diretamente o aspecto *duração da bateria*. Já uma opinião *indireta* é expressada indiretamente, baseando-se nos efeitos que a entidade avaliada causa em outra entidade. Esse

tipo de opinião é frequentemente encontrado na área médica [Liu 2012] – a frase "*após tomar o medicamento, senti fortes dores no estômago*", por exemplo, não expressa nenhuma opinião direta sobre a entidade medicamento, porém pode-se indiretamente inferir um sentimento negativo devido às dores causadas.

- **Opiniões explícitas e implícitas:** uma opinião *explícita* é uma sentença subjetiva da qual se pode diretamente extrair uma opinião, por exemplo: "*Windows é um bom sistema operacional*" ou "*Windows é melhor que Linux*". Uma opinião *implícita* é uma sentença objetiva que sugere uma opinião. Esse tipo de sentença geralmente expressa fatos, por exemplo: "*comprei a roupa há uma semana e o tecido começou a soltar fios*" e "*os notebooks da Asus duram mais que os da Dell*".

Opiniões diretas e explícitas são mais fáceis de lidar, portanto a maior parte da pesquisa na área é concentrada nelas. Opiniões implícitas são difíceis de serem detectadas, pois dependem muito mais do resultado semântico da sentença que as explícitas e, desse modo, são mais complexas do ponto de vista de PLN. A dificuldade principal deve-se ao fato de que esse tipo de opinião é altamente relacionado ao domínio e ao contexto em que se encontram [Zhang e Liu 2011]. Uma frase que indica uma opinião implícita em uma *review* sobre um produto eletrônico possivelmente apresenta outro significado quando encontrada em uma *review* sobre filmes – portanto é difícil aplicar técnicas genericamente.

2.2.3 Tarefas da Análise de Sentimentos

Com uma definição de opinião bem estabelecida, é possível apresentar as principais tarefas do processo de mineração de opinião. Segundo [Liu 2012], o objetivo da análise de sentimentos é encontrar todas as quíntuplas de opinião $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ em um dado documento de opinião d. Encontrar cada um dos componentes da tupla pode ser considerado uma tarefa diferente.

A primeira tarefa é encontrar a entidade principal e sobre qual a opinião trata. Esse processo se assemelha ao reconhecimento de entidades mencionadas, mais conhecido pelo termo em inglês *named-entity recognition* (NER). NER é uma tarefa de extração de informação que tem como objetivo classificar elementos do texto em categorias pré-estabelecidas, como pessoas, organizações, produtos etc. Esse processo pode ser realizado por meio de técnicas não-supervisionadas e já existem algoritmos bem consolidados que extraem entidades e obtém re-

sultados com mais de 94% de precisão, como o proposto por [Collins e Singer 1999]. Existem também adaptações de algoritmos para extrair entidades especificamente de redes sociais, como o trabalho de [Ritter et al. 2011], que adapta técnicas básicas de PLN para trabalhar adequadamente com o tipo de linguagem utilizada em *tweets*. Em avaliações de consumidores sobre produtos, a entidade e geralmente é explícita e pode ser extraída diretamente – pois a avaliação trata de um produto específico.

Cada aspecto a da entidade também precisa ser identificado. Por exemplo, na frase "*a qualidade das fotos dessa câmera é incrível*", "*qualidade das fotos*" é um aspecto que deve ser extraído. Nesse caso, não é possível utilizar técnicas de NER, pois aspectos não podem ser classificados em categorias pré-definidas por variarem muito dependendo do contexto. Aspectos encontradas em opiniões sobre produtos são bem diferentes de os encontrados em opiniões políticas, desse modo algoritmos de NER não são adequados. Para extração de aspectos são empregadas técnicas que utilizam a relação semântica entre as palavras que compõe a opinião. Em geral, aspectos são compostos por substantivos e frequentemente estão em frases nominais – essa é a ideia básica de técnicas como a proposta em [Hu e Liu 2004], que serão discutidas mais adiante.

O terceiro componente é o sentimento s. Classificá-lo como positivo, negativo ou neutro constitui uma tarefa. Técnicas que utilizam bases lexicais são comumente empregas na classificação de sentimentos. A abordagem proposta por [Kim e Hovy 2004] consiste basicamente em reunir algumas palavras que expressam sentimentos, que foram manualmente classificadas em positivas e negativas, de modo a formar uma base. Essa base é então alimentada utilizando bases lexicais como [University 2015], que possuem listas de sinônimos. O objetivo é encontrar sinônimos das palavras já registradas na base e adicioná-los, classificando-os com a mesma polaridade da palavra-base.

Os dois últimos componentes são a fonte da opinião h e a data t. Eles também precisam ser extraídos e categorizados. A fonte da opinião pode ser a pessoa ou a organização que expressou a opinião – nesse casos, técnicas de NER podem ser utilizadas. Em avaliações de produtos encontradas em *websites* e em *blogs*, o autor da postagem é geralmente a fonte da opinião, podendo ser diretamente extraído. No entanto, em opiniões veiculadas em redes sociais isso pode não ser aplicável – pois é comum o compartilhamento de opiniões de terceiros. A

extração da data segue os mesmos princípios, sendo que a data de publicação de uma opinião não é necessariamente a mesma data de sua concepção. Portanto, é preciso definir claramente o significado desse atributo.

Resumindo as tarefas descritas acima, dado um conjunto de documentos de opinião D, a análise de sentimentos consiste em seguir as seis tarefas [Liu 2012]:

Tarefa 1: extrair todas as entidades contidas em D e categorizá-las em grupos de entidades sinônimas. Cada grupo representa uma entidade e_i.

Tarefa 2: extrair todos os aspectos das entidades e categorizá-los em grupos. Cada grupo representada um aspecto a_{ij}.

Tarefa 3: extrair todas as fontes de opinião do texto e categorizá-las de forma análoga às tarefas acima.

Tarefa 4: extrair todas as datas em que as opiniões do texto foram expressadas e armazená-las em um formato padrão (por exemplo *dd/mm/aaaa*).

Tarefa 5: para cada aspecto a_{ij}, determinar qual o sentimento relacionado a ele e classificá-lo como positivo, negativo ou neutro, ou ainda o atribuir um valor numérico.

Tarefa 6: gerar todas as quíntuplas de opinião $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ de cada documento de opinião $d \in D$.

O exemplo abaixo, adaptado de [Liu 2012], ilustra o processo de análise de sentimentos:

Escrito por: João **Data: 06/09/2015**

"(1) Comprei uma câmera da Samsung semana passada e meu amigo comprou uma da Canon. (2) Nos últimos dias, temos usado bastante nossas câmeras. (3) As fotos da minha Samy não são tão boas e duração da bateria é curta. (4) Meu amigo está bem feliz com a câmera dele e está amando a qualidade das fotos. (5) Quero uma câmera que tire fotos boas também. (6). Vou devolvê-la na loja amanhã."

Figura 2.6: Exemplo de uma avaliação de consumidor sobre uma câmera fotográfica

A **Tarefa 1** deve extrair as entidades "Samsung", "Canon"e "Samy" e agrupar "Samsung" e "Samy", pois representam a mesma entidade. A **Tarefa 2** deve extrair os aspectos "fotos",

"qualidade das fotos" e "duração da bateria"e agrupar "fotos"e "qualidade das fotos", pois representam a mesma característica. A **Tarefa 3** deve identificar João (autor da avaliação) como fonte da opinião da frase (3) e seu amigo como fonte da (4). A **Tarefa 5** deve identificar que a frase (3) expressa opiniões negativas sobre a qualidade das fotos e a duração da bateria da câmera Samsung e que a frase (4) expressa opinião positiva sobre a câmera Canon em geral e sobre sua qualidade das fotos. Por fim, a **Tarefa 6** deve gerar as quíntuplas de opinião, que podem ser representadas como:

(Samsung, qualidade_fotos, negativo, João, 06/09/2015)
(Samsung, duração_bateria, negativo, João, 06/09/2015)
(Canon, Canon, positivo, amigo, 06/09/2015)
(Canon, qualidade_fotos, positivo, amigo, 06/09/2015)

Como a terceira opinião refere-se à entidade como um todo, é possível representar o aspecto pelo próprio nome da entidade. Também é comum utilizar o termo "GERAL", que indica que a opinião trata sobre o a entidade como um todo.

2.2.4 Dificuldades na Análise de Sentimentos

Opiniões são escritas em linguagem natural de forma não-estruturada. Apesar dos avanços na área de PLN, é difícil representar computacionalmente esse tipo de informação. Nesse contexto, existe um conceito importante relacionado a opiniões e sentimentos – o conceito de *subjetividade*. Uma *sentença objetiva* apresenta informação factual a respeito de uma entidade, enquanto uma *sentença subjetiva* expressa um ponto de vista pessoal. Expressões subjetivas são encontradas na forma de opiniões, reclamações, alegações, acusações, especulações, desejos e crenças [Riloff e Wiebe 2003]. Opiniões, portanto, são subjetivas. Porém nem toda sentença subjetiva expressa uma opinião. Determinar se uma frase contém opinião é um desafio da análise de sentimentos, sendo o foco de estudo da subárea de pesquisa conhecida como *classificação de subjetividade*.

O principal objetivo dessa subárea é desenvolver classificadores capazes de distinguir sentenças subjetivas de opinativas [Wiebe e Riloff 2005]. A frase (5) na Figura 2.6 é subjetiva, visto que expressa um desejo, porém não contém opinião alguma. Por outro lado, frases objetivas podem indiretamente expressar opinião. A frase "*a estampa desbotou após a primeira lavagem*"

não é subjetiva, pois é apenas uma constatação. No entanto, ela sugere uma opinião negativa sobre a qualidade da estampa – visto que desbotar é um efeito indesejado. Considerando isso, é ainda mais difícil determinar se uma sentença é opinativa ou não.

Outra dificuldade da análise de sentimentos é a diferença de pontos de vista. Uma característica considerada negativa pelo autor da opinião pode ser considerada de maneira diferente pelo leitor. A frase "*o preço dos imóveis caiu e isso é ruim para a economia*" expressa um sentimento negativo no ponto de vista de seu autor. Porém ao considerar o ponto de vista de um potencial comprador de imóveis, o sentimento é positivo. Uma possível solução é ignorar o problema e considerar apenas o ponto de vista do autor.

Há também dificuldades inerentes à linguagem natural – como a polissemia (multiplicidade de sentidos de uma palavra). Um problema frequente é que a mesma palavra pode indicar polaridades de sentimento diferentes dependendo do contexto. Na frase "*o chuveiro esquenta bem a água mesmo no inverno*", a palavra "esquenta" sugere opinião positiva. Já na frase "*o computador esquenta bastante mesmo no inverno*" a palavra indica opinião negativa, visto que nesse contexto esquentar é um efeito indesejado. Como a classificação de sentimentos utiliza frequentemente a abordagem baseada em sinônimos, contornar essa dificuldade é uma tarefa árdua. Frases sarcásticas também são difíceis de lidar, porém são raras em avaliações de produtos. A frase "*que aspirador ótimo! Funcionou duas vezes...*" contém uma palavra que indica sentimento positivo ("ótimo"), porém seu significado é justamente o oposto. Esse tipo de sentença é oriunda de recursos avançados da linguagem natural – cuja abstração computacional é difícil.

A definição desses conceitos é necessária para poder representá-los computacionalmente e desenvolver técnicas. No próximo capítulo, o processo de mineração de opinião será apresentado do ponto de vista técnico.

Capítulo 3

O Processo de Mineração de Opinião

No Capítulo 2, as tarefas da análise de sentimentos foram apresentadas conceitualmente. Nesse capítulo, elas serão discutidas do ponto de vista técnico e computacional. Na taxonomia proposta por [Liu 2012], a análise de sentimentos pode ser efetuada em três níveis: nível de documento, nível de sentença e nível de entidade-aspecto. Cada nível aborda a análise com granularidade diferente, sendo:

- **Nível de Documento:** o objetivo desse nível de análise é classificar a opinião expressada em um documento como um todo. No caso de avaliações de consumidores sobre produtos, esse nível classifica se a avaliação no geral expressa um sentimento negativo ou positivo sobre o produto. Considera-se que o documento apresenta opiniões referentes apenas a uma entidade (um único produto, por exemplo), portanto, esse nível não é adequado para tratar de opiniões comparativas ou que consideram diversas entidades.

- **Nível de Sentença:** esse nível analisa todas as sentenças contidas em um documento de opinião. O objetivo é classificá-las individualmente como positivas, negativas ou neutras (que indicam que não há opinião expressada). Esse nível de análise se assemelha ao problema de classificação de subjetividade, apresentado na Seção 2.2.4.

- **Nível de Entidade-Aspecto:** é o nível que realiza a análise com granularidade mais fina. Em vez de tentar identificar opiniões em estruturas linguísticas (como parágrafos, sentenças e documentos), o nível de entidade-aspecto analisa diretamente as opiniões pra então identificar ao que elas estão associadas – ou seja, considera que opiniões são formadas por alvos e por sentimentos. Assim, é possível encontrar no documento de opinião diferentes entidades e aspectos dessas entidades, bem como os sentimentos relacionados a eles.

O nível de documento apresenta a análise mais genérica e é o mais limitado, por considerar apenas uma entidade. É adequado para documentos de opinião concisos, nos quais a entidade avaliada é única e explícita. O nível de sentença é um pouco mais específico e analisa todas as sentenças do documentos individualmente. No entanto, apresenta os mesmos problemas do nível de documento, pois considera apenas a sentença em si e não leva em conta a relação entre todas as sentenças que compõe o documento. Por fim, o nível de entidade-aspecto realiza a análise mais detalhada, identificando todas as entidades e os aspectos que são avaliados no documento de opinião. É, porém, o nível mais complexo devido à profundidade da análise. Nas seções seguintes, serão brevemente discutidas técnicas para efetuar cada um dos níveis de análise.

3.1 Classificação de Sentimentos a Nível de Documento

A classificação de sentimentos é um dos tópicos mais estudados da área [Pang e Lee 2008]. Seu objetivo é classificar um documento de opinião quanto à opinião expressada por ele, seja positiva ou negativa. Ou seja, dado um documento de opinião d que avalia uma entidade e, o objetivo é determinar o sentimento geral s dado pela fonte da opinião sobre a entidade, isto é, determinar a quíntupla $(\_, GERAL, s, \_, \_)$. A entidade e, a fonte da opinião h e a data t são consideradas disponíveis diretamente no documento ou irrelevantes.

Essa classificação considera que todo o documento trata sobre apenas uma entidade e apresenta a opinião de uma única fonte. Na prática, essa suposição pode acarretar em perda de informação. O autor de uma avaliação sobre um produto pode expressar opinião positiva sobre determinadas características e opinião negativa sobre outras. A classificação em nível de documento não leva em conta essa diferença e efetua a classificação baseando-se apenas no resultado geral das opiniões encontradas.

Na próxima seção, serão brevemente apresentadas técnicas que visam solucionar o problema da classificação de sentimentos a nível de documento.

3.1.1 Classificação por Técnicas Supervisionadas

A classificação de sentimentos pode ser formulada de maneira semelhante a um problema de classificação da área de mineração de dados (MD). Na MD, classificação refere-se à tarefa

de prever a classe que determinado valor pertence [Zaki e Jr 2014]. As classes indicam as categorias nas quais os valores podem ser classificados. No contexto de sentimentos, as classes são *positivo* e *negativo*. A maioria das aplicações não considera a classe *neutro*, visto que ela representa ausência de opinião. Técnicas supervisionadas indicam que há uma etapa de treinamento, a qual consiste em analisar dados já pré-classificados e por meio disso prever a classe dos demais dados. Esse processo é geralmente efetuado utilizando avaliações de consumidores sobre produtos, pois estas já possuem uma classe explícita dada pelo autor da opinião, usualmente indicada por uma avaliação numérica (como número de estrelas).

Os métodos tradicionais de classificação, como classificadores Bayesianos e *support vector machines* (SVM), podem ser empregados na classificação de sentimentos. Classificadores Bayesianos utilizam diretamente o teorema de Bayes para prever a classe de uma nova instância x. A probabilidade a posteriori $P(c_i|x)$ de cada classe c_i é estimada e a com maior probabilidade é escolhida. Segundo [Zaki e Jr 2014], a classe prevista para x é dada por:

$$\hat{y} = \arg\max\{P(c_i|x)\} \tag{3.1}$$

O teorema de Bayes permite que a Equação 3.1 possa ser reescrita em termos da verossimilhança e da probabilidade a priori, resultando na Equação 3.2.

$$P(c_i|x) = \frac{P(x|c_i) \cdot P(c_i)}{P(x)} \tag{3.2}$$

Sendo $P(x|c_i)$ a verossimilhança, definida como a probabilidade de observar x assumindo que sua classe é c_i, $P(c_i)$ a probabilidade a priori da classe c_i e $P(x)$ a probabilidade de observar x em qualquer classe, também chamada de evidência. As probabilidades são estimadas utilizando modelos de distribuição, como o Gaussiano, o multinomial e o de Bernoulli. Na classificação textual, a distribuição de Bernoulli é frequentemente utilizada [McCallum e Nigam 1998]. Esse tipo de classificador é eficiente quando os atributos são independentes da classe, porém o trabalho de [Domingos e Pazzani 1997] mostra que resultados satisfatórios podem ser obtidos mesmo em atributos dependentes.

Support vector machines, ou máquinas de vetores suporte, são classificadores baseados na ideia de encontrar uma região que permita separar as classes. Seu objetivo principal é separar os dados de entrada em hiperplanos, que indicam a superfície de decisão [Berwick 2015]. Os métodos matemáticos que constituem o funcionamento das SVMs são complexos e fogem

do escopo desse trabalho. Trabalhos como [Cristianini e Shawe-Taylor 2000] e [Burges 1998] apresentam detalhadamente todo o processo de classificação das SVMs.

Em [Pang, Lee e Vaithyanathan 2002], são utilizados classificadores Bayesianos e SVMs para classificar avaliações sobre filmes como positivas ou negativas. Diferente da tarefa de classificação da MD tradicional, a classificação textual não trabalha com dados numéricos – mas sim com palavras, que por sua vez não tem significado computacional direto. Portanto, é necessário definir alguma característica (ou *feature*) da palavra que possa ser quantizada, para então aplicar os métodos dos classificadores [Joachims 1998]. No artigo citado, unigramas são utilizadas como *features* para classificação. Em PLN, um **n-grama** é uma sequência de **n** palavras utilizada para computar a probabilidade de determinada sentença ser formada [Jurafsky e Martin 2009]. Um n-grama de tamanho 1 é chamado de unigrama. O modelo de unigrama indica a probabilidade de cada palavra aparecer no texto. Por considerarem probabilidade, n-gramas são características viáveis para serem utilizadas na classificação textual.

3.1.2 Classificação por Técnicas Não-Supervisionadas

Diferente da tarefa de classificação de MD, a classificação de sentimentos pode ser feita por técnicas não-supervisionados, pois o processo de classificação refere-se a encontrar a polaridade do sentimento – e não propriamente separá-lo em classes, como na MD. Portanto, é possível empregar técnicas que não requerem treinamento nem dados previamente classificados. [Turney 2002] propõe uma técnica para classificar a polaridade de avaliações de consumidores utilizando a orientação semântica das frases da avaliação que contém adjetivos ou advérbios. Palavras com essas classes gramaticais indicam sentimento. A orientação semântica é calculada como a medida de informação mútua entre determinada frase e a palavra "*excellent*" menos a informação mútua entre a frase e "*poor*". Segundo o autor, a técnica atinge em média uma taxa de 74% de acerto. O algoritmo proposto consiste em três passos:

Passo 1: duas palavras consecutivas são extraídas se seguirem o padrão indicado na Tabela 3.1, adaptada de [Turney 2002].

A razão para o padrão usado é que adjetivos e advérbios geralmente indicam palavras que expressam opinião.

Passo 2: a orientação semântica das frases extraídas é calculada pela informação mútua.

Tabela 3.1: Padrão semântico das palavras a serem extraídas

Primeira Palavra	**Segunda Palavra**	**Terceira Palavra** (não extraída)
Adjetivo	Substantivo	Qualquer
Advérbio	Adjetivo	não-Substantivo
Adjetivo	Adjetivo	não-Substantivo
Substantivo	Adjetivo	não-Substantivo
Advérbio	Verbo	Qualquer

Uma das métricas para calculá-la é a *pointwise mutual information* (PMI), descrita na Equação 3.3, que tem como objetivo mensurar a dependência estatística entre dois termos.

$$PMI(termo1, termo2) = \log_2 \left(\frac{Pr(termo1 \wedge termo2)}{Pr(termo1) \cdot Pr(termo2)} \right) \tag{3.3}$$

Na Equação 3.3, $Pr(termo1 \wedge termo2)$ indica a probabilidade de ocorrência consecutiva dos termos e $Pr(termo1) \cdot Pr(termo2)$ indica a probabilidade de ocorrência consecutiva quando os termos são estatisticamente independentes – ou seja, não tem relação semântica. A orientação semântica (OS) é por fim calculada baseando-se nas palavras "*excellent*" e "*poor*", como na Equação 3.4. Segundo o autor, essas palavras foram escolhidas devido ao fato de que em sistemas de avaliação que consideram número de estrelas, é comum definir avaliações de uma estrela como "*poor*"e de cinco estrelas como "*excellent*". Assim:

$$OS(frase) = PMI(frase, "excellent") - PMI(frase, "poor") \tag{3.4}$$

As probabilidades são estimadas por meio de consultas a motores de busca. Como exemplo, considere a frase "*taxas reduzidas*", que se enquadra no padrão da Tabela 3.1. Para calcular sua probabilidade, ela é consultada por si só em um motor de busca, bem como seguida por "*excellent*" e "*poor*". A quantia de resultados retornados para cada frase permite a estimativa da probabilidade.

Passo 3: dada uma avaliação de consumidor, o algoritmo computa a orientação semântica média de todas as frases extraídas e classifica a avaliação como positiva caso a média seja positiva e como negativa caso contrário.

Esse algoritmo apresenta resultados satisfatórios, com a taxa de acerto variando de 66% para avaliações sobre filmes a 84% para avaliações sobre automóveis.

3.2 Classificação de Sentimentos a Nível de Sentença

Como discutido, a análise a nível de documento é muito genérica e pode não ser adequada para muitas aplicações. Nessa seção, o nível de sentença será apresentado – ou seja, o nível que classifica o sentimento expressado em cada sentença. É importante ressaltar que fundamentalmente não há diferença entre as classificações de documento e de sentença, pois sentenças são basicamente documentos curtos. No entanto, o problema é abordado de maneira diferente nesse nível. Dada uma sentença x, o objetivo é determinar se x expressa uma opinião positiva, negativa ou neutra (sem opinião).

A quíntupla de opinião não é utilizada nesse nível, pois apenas sentenças são consideradas e, portanto, não é possível identificar entidades e aspectos. Apesar de ser de uso limitado, a classificação em nível de sentença constitui um passo intermediário para outras aplicações e é útil em muitos casos. Ela pode ser definida como dois problemas diferentes. O primeiro, chamado de *classificação de subjetividade*, é determinar se dada sentença apresenta opinião ou não. O segundo classifica sentenças opinativas em positivas ou negativas. A análise em nível de sentença será discutida do ponto de vista dos dois problemas.

3.2.1 Classificação de Subjetividade

A classificação de subjetividade divide as sentenças em duas classes: subjetivas e objetivas. Uma sentença objetiva expressa informações factuais e constatações, enquanto uma subjetiva expressa pontos de vista pessoais e opiniões. Como visto na seção 2.2.4, sentenças objetivas podem ser encontradas de várias formas. Inicialmente, a classificação de subjetividade surgiu como um problema isolado, sem considerar a classificação de sentimentos. Em pesquisas mais recentes ela é tratada como o passo inicial para classificar sentimentos, para pós isso efetuar tarefas mais complexas, como extração de aspectos [Liu 2012].

As técnicas para realizar essa tarefa são em maioria supervisionadas. [Wiebe, Bruce e O'Hara 1999] propõe uma abordagem utilizando classificadores Bayesianos e as seguintes *features* para classificação: *features* binárias para indicar a presença na sentença de um pronome, um adjetivo, um número cardinal, um modal e um advérbio; *feature* binária representando se a sentença é início de parágrafo e uma *feature* binária representando a ocorrência consecutiva de palavras classificadas como subjetivas e objetivas. O autor sugere o

uso de *features* mais complexas para alcançar resultados mais precisos.

Técnicas não-supervisionadas também podem ser utilizadas na classificação de subjetividade. [Wiebe 2000] propõe uma abordagem que utiliza a presença de expressões subjetivas na sentença para classificá-la. Um conjunto de expressões subjetivas é usado como base e distribuições estatísticas são utilizadas para expandir a base e encontrar novas palavras que indicam subjetividade. Esse conjunto de palavras é então usado para classificar a sentença. Também são aplicados filtros que visam corrigir sentenças classificadas erroneamente.

Abordagens não-supervisionadas mais modernas utilizam métodos baseados em regras que indicam subjetividade e em padrões de sentenças já classificadas como subjetivas. O algoritmo proposto em [Wiebe e Riloff 2005] primeiramente classifica uma sentença como subjetiva se ela contém dois ou mais indicativos de subjetividade fortes, baseando-se em regras. Segundo os autores, as regras apresentam indícios bem estabelecidos sobre subjetividade, porém elas não apresentadas no texto. Em seguida, são extraídos padrões das sentenças classificadas para estender o conjunto de regras.

3.2.2 Classificação de Sentimentos

Se uma sentença é classificada como subjetiva, ela pode então ser classificado como positiva ou negativa. Os métodos apresentadas na seção 3.1 podem ser aplicadas às sentenças individuais, visto que elas são por si só documentos de opinião. A classificação de sentimento em nível de sentença assume que uma sentença expressa uma única opinião advinda de uma única fonte. Essa suposição é válida para sentenças simples como "*o desempenho desse notebook é ótimo*". Porém não é para sentenças complexas e compostas como "*o desempenho desse notebook é ótimo, mas o design e o peso deixam a desejar*". Nesse exemplo, a sentença expressa sentimento positivo sobre o aspecto "*desempenho*"e negativo sobre "*design*"e "*peso*". Considerá-la como um todo não extrai esses diferentes sentimentos.

A técnica para classificação de sentimentos em sentenças subjetivas proposta por [Yu e Hatzivassiloglou 2003] utiliza uma abordagem similar à de [Turney 2002], exposta na Seção 3.1.2. Essa técnica usa uma grande base de adjetivos como parâmetro de classificação, diferente das apenas duas palavras ("*excellent*" e "*poor*") usadas em [Turney 2002]. Além disso, outra métrica estatística é utilizada no lugar da PMI para avaliar a polaridade de cada adjetivo,

advérbio, substantivo e verbo.

O trabalho de [Hassan, Qazvinian e Radev 2010] apresenta um método para identificar a atitude de participantes de fóruns de discussão *online* em relação um ao outro. O primeiro passo do algoritmo proposto extrai sentenças com pronomes em segunda pessoa, pois essa classe de palavra indica que o autor da sentença se refere a outro usuário. Em seguida, a polaridade de cada palavra da sentença é identificada. O próximo passo é extrair padrões que generalizem as sequências. Esses padrões são utilizados para criar dois modelos ocultos de Markov [Stamp 2015] para cada padrão. O primeiro modelo caracteriza a relação entre diferentes palavras que correspondem a sentenças que indicam atitude. O segundo é análogo ao primeiro, porém para sentenças que não indicam atitude. Dada uma nova sentença, o algoritmo estima a probabilidade de essa sentença ser gerada por cada um desses modelos e a estimativa é utilizada para definir sua polaridade.

3.3 Análise em Nível de Entidade-Aspecto

Como discutido, classificar opiniões em nível de documento e de sentença não é suficiente para a maioria das aplicações. Para uma análise mais completa, é necessário identificar os diferentes aspectos da entidade e o sentimento expressado por cada um. Nesse nível, o objetivo é extrair toda a tupla de opinião $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$. Para exemplificar, consideram-se opiniões expostas em avaliações de consumidores sobre produtos. Nesse caso, supõe-se que a entidade avaliada e, a fonte da opinião h e a data t estão disponíveis diretamente – isto é, não precisam ser extraídos. Portanto, a tarefa se resume a extrair os aspectos e o sentimento relacionado a eles.

Extração de aspectos: essa tarefa visa extrair os aspectos que estão sendo avaliados. Por exemplo, na frase "*a qualidade de vídeo dessa webcam é decepcionante*", o aspecto "*qualidade de vídeo*" da entidade "*webcam*" deve ser extraído. É importante ressaltar que é fundamental saber a qual entidade o aspecto pertence, pois é a relação entidade-aspecto que permite identificar apropriadamente as opiniões.

Classificação de sentimentos de aspectos: essa tarefa determina se as opiniões relacionadas aos aspectos são positivas ou negativas. No exemplo acima, a opinião sobre o

aspecto "*qualidade de vídeo*" é positiva. É também possível classificar o sentimento geral como positivo.

A classificação de sentimentos funciona da mesma maneira que a já apresentada nos outros níveis. Já a extração de aspectos é abordada por um conjunto próprio de técnicas. Algumas dessas técnicas serão apresentadas nas próximas seções.

3.3.1 Extração de Aspectos

A extração de aspectos pode ser abordada como uma tarefa de extração de informação [Liu 2012]. No contexto da análise de sentimentos, há características específicas que definem o processo de extração. Um ponto chave é que uma opinião sempre tem um alvo – o qual frequentemente indica um aspecto a ser extraído. É importante ressaltar que algumas expressões de opinião podem, além de indicar sentimento, implicitamente conter aspectos. Por exemplo, na frase "esse sofá é caro" a palavra "*caro*" indica um sentimento negativo, mas também se refere implicitamente ao aspecto *preço* da entidade. As técnicas que serão apresentadas são focadas em extrair aspectos explícitos. Há quatro abordagens principais:

1. Extração utilizando modelos de tópico;
2. Extração utilizando aprendizado supervisionado;
3. Extração baseada em substantivos frequentes e frases nominais;
4. Extração baseada na relação entre opinião e alvo;

O foco desse trabalho é na terceira abordagem, que será discutida com mais detalhes. As demais serão brevemente apresentadas e a literatura adequada para aprofundar-se nas técnicas será indicada.

3.3.1.1 Extração por Modelos de Tópico

Modelos de tópico são um conjunto de algoritmos cujo objetivo é descobrir a estrutura temática de documentos textuais [Blei 2015]. Os modelos consideram que documentos são compostos de temas (também chamados assuntos ou tópicos) e que cada tema é uma distribuição probabilística de palavras. Os algoritmos agrupam as palavras do texto em *clusters* que indicam

o tópico a qual pertencem. Há dois modelos principais: o *probabilistic latent semantic analysis* (pLSA), proposto por [Hofmann 1999], e o *latent Dirichlet allocation* (LDA), proposto por [Blei, Ng e Jordan 2003]. A Figura 3.1, adaptada de [Blei 2015], ilustra a intuição do modelo LDA.

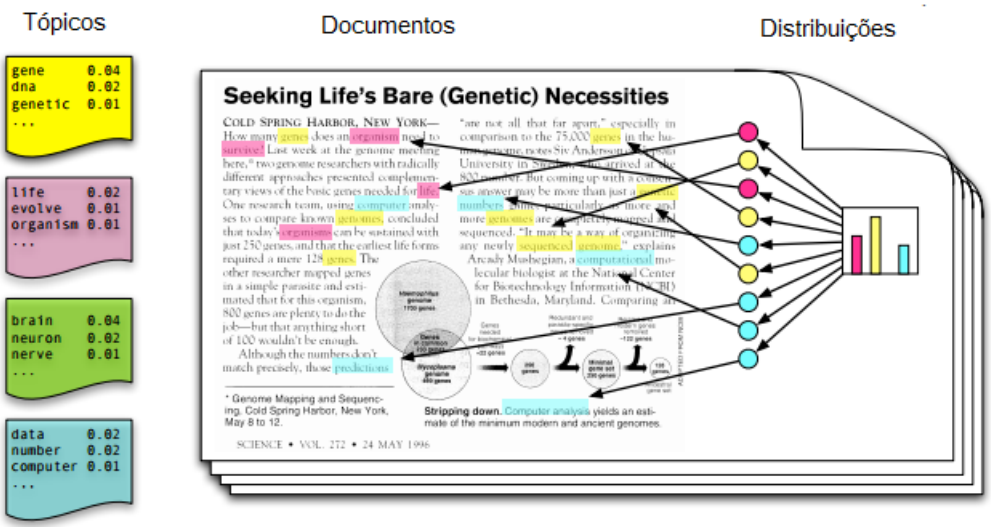

Figura 3.1: Ilustração do modelo LDA

Assume-se que existe determinado número de tópicos para os documentos, ilustrados à esquerda da Figura 3.1. Considera-se que cada documento é gerado da seguinte maneira: uma distribuição sobre os tópicos (o histograma à direita) é selecionada. Em seguida, um tópico é atribuído para cada palavra (os círculos). Uma palavra correspondente ao tópico é então destacada no texto.

No contexto da análise de sentimentos, tópicos podem ser considerados aspectos e sentimentos. Porém, é necessário extraí-los separadamente. Essa separação pode ser realizada estendendo modelos de tópico básicos. [Lin e He 2009] propõe uma extensão do LDA que é adequada para análise de sentimentos, pois além da modelagem de tópicos, utiliza modelos para considerar sentimentos e os inclui no cálculo da distribuição.

3.3.1.2 Extração por Aprendizado Supervisionado

A extração de aspectos pode ser considerada uma caso específico do problema de extração de informação. Há muitos métodos consolidados para extração de informação, como [Mooney e Bunescu 2005]. Por serem métodos supervisionados, necessitam de uma base de

dados previamente classificada para realizar o processo de treinamento. Técnicas já citadas anteriormente, como SVMs e HMMs, podem também ser utilizadas nesse contexto. Em [Jin e Ho 2009], é apresentado um algoritmo que aplica HMMs para aprender padrões e extrair aspectos e expressões de opinião. Já [Manevitz e Yousef 2002] propõe uma técnica parcialmente supervisionada para extrair aspectos utilizando uma versão adaptada de SVMs de classe única – que visam identificar se o padrão pertence ou não à determinada classe. Na adaptação proposta, dados discrepantes representam uma segunda classe.

3.3.1.3 Extração Baseada em Substantivos Frequentes e Frases Nominais

Essa abordagem extrai aspectos frequentes, que são substantivos e frases nominais que constam repetidas vezes no documento de opinião. Uma frase nominal é a frase construída sem verbos e que pode exercer o papel um substantivo. O algoritmo proposto por [Hu e Liu 2004] é baseado nessa ideia. Primeiramente, são extraídas todas as palavras que são substantivos ou as frases que são nominais. O conjunto extraído representa o total de termos que são candidatos a aspectos. Novos candidatos são gerados concatenando pares e trios de termos que aparecem na mesma sentença. Por exemplo, se *duração da bateria* e *desempenho* estão na mesma sentença, o termo *duração da bateria desempenho* é adicionado ao conjunto. Em seguida, uma medida chamada *p-suporte* é calculada para cada candidato. O *p-suporte* de um termo t é o número de sentenças que contém t, excluindo as que contém um termo t' que engloba t. Por exemplo, a frase "*a duração da bateria é boa*" só é contada para o *p-suporte* de "*duração da bateria*", mas não para o de "*bateria*", pois um engloba o outro. Após o cálculo do *p-suporte*, o algoritmo utiliza métodos de poda para remover candidatos inválidos.

O primeiro passo de poda remove candidatos compostos por várias palavras que são não-compactos em mais de uma sentença. Um termo t é dito não-compacto em uma sentença s se a distância em s entre as palavras que compõem t é maior que 3. Por exemplo, na frase "*a duração da bateria é muito melhor que a tela*" o termo candidato "*duração da bateria tela*" é não-compacto. O segundo passo remove termos t com *p-suporte* menor que 3 que estão contidos em um termo maior t'. Por fim, os candidatos com *p-suporte* maior que um limiar são extraídos e considerados aspectos. O valor do limiar é definido empiricamente e depende da aplicação.

3.3.1.4 Extração por Relação entre Opinião e Alvo

Considerando que opiniões possuem alvos, há uma relação clara entre eles. Essa relação pode ser explorada para extrair aspectos que não são frequentes no documento de opinião. A ideia dessa abordagem é simples. Considera-se, inicialmente, que os aspectos frequentes já foram extraídos. Se alguma sentença possui palavras que indicam sentimentos porém não contém um aspecto frequente, é provável que ela ainda contenha um aspecto que não foi extraído. O substantivo ou a frase nominal mais próximo do sentimento é então extraído e considerado como um aspecto. Por exemplo, na frase "*o sistema operacional é eficiente*" sabe-se que "*eficiente*" indica sentimento, portanto a frase nominal "*sistema operacional*" é extraída como aspecto. Apesar de simples, essa técnica funciona bem mesmo quando aplicada por si só [Liu 2012]. A técnica proposta em [Blair-Goldensohn et al. 2008] utiliza padrões de sentimentos e segue uma uma ideia semelhante.

3.4 Trabalhos Correlatos

Este trabalho pode ser avaliado de duas formas: como uma introdução conceitual à área de análise de sentimentos e como o desenvolvimento de um sistema de mineração de opinião em nível entidade-aspecto. Portanto, serão apresentados trabalhos correlatos referentes a essas duas vertentes.

Do ponto de vista de introdução conceitual, o trabalho de [Liu 2012] é a maior referência. O livro apresenta um panorama geral sobre a área, discutindo sobre os principais tópicos de pesquisa com grande abrangência. Seu objetivo não é aprofundar-se em nenhum tópico específico, mas a literatura pertinente ao contexto é sempre apresentada. Há mais de 400 referências de artigos. Além disso, o autor do livro também é autor de trabalhos fundamentais na área. A tese de [Bross 2013] trata de análise de sentimentos orientada a aspectos, porém também apresenta uma revisão bibliográfica completa sobre a área, discutindo diversos conceitos importantes. O trabalho de [Pang e Lee 2008] é menos abrangente que o de [Liu 2012], mas se aprofunda mais no ponto de vista técnico.

Do ponto de vista de sistema, o trabalho de [Hu e Liu 2004] é a base do método que é implementado. Em [Pavlopoulos e Androutsopoulos 2014], são apresentadas melhorias ao método de [Hu e Liu 2004], que foram consideradas no sistema. Nesse trabalho os algoritmos são

discutidos de forma mais clara – facilitando assim a reprodutibilidade. [Jo e Oh 2011] trata sobre extração de aspectos em avaliações de consumidores sobre produtos utilizando modelos de tópico. Apesar de as técnicas utilizadas serem diferentes, a proposta é semelhante à desse trabalho, inclusive utilizando o mesmo tipo de documento de opinião. A base de dados utilizada na avaliação experimental está disponível publicamente para *download*, portanto poderá ser usada como critério de avaliação do sistema proposto. O trabalho de [Bagheri, Saraee e Jong 2013] também apresenta métodos para extração de aspectos em avaliações sobre produtos. A técnica utilizada tem conceitos em comum com a empregada no sistema proposto. As regras usadas para identificação de aspectos candidatos são mais complexas, e desse modo podem atuar no refinamento do algoritmo implementado pelo sistema.

Capítulo 4

O Sistema Proposto

Além de apresentar uma introdução à área de mineração de opinião, um sistema computacional capaz de efetuar análise de sentimentos em nível de entidade-aspecto foi implementado, com objetivo de extrair opiniões contidas em avaliações de consumidores (*reviews*) sobre produtos. Considerando que muitas vezes existem milhares de avaliações sobre o mesmo produto, a tarefa é gerar automaticamente resumos contendo as opiniões identificadas para os diferentes aspectos do produto. Esses resumos visam facilitar a compreensão das informações por parte dos clientes, assim como dos fabricantes.

O sistema foi desenvolvido na linguagem de programação Python em conjunto à plataforma *Natural Language Toolkit* (NLTK) [Project 2015]. A escolha da linguagem se deve ao fato que Python é simples e possui funcionalidades para processar dados linguísticos [Bird, Klein e Loper 2009]. Outro ponto importante é a plataforma NLTK, um conjunto de bibliotecas que provê diversos métodos utilizados no processamento de texto, como *tokenizers*, *stemmers*, *taggers* e *parsers*. É importante ressaltar que o sistema é desenvolvido para processar *reviews* em língua inglesa. Todos os algoritmos que implementam técnicas de PLN empregados em sua construção são próprios para essa língua.

De modo geral, a função do sistema proposto é gerar um resumo das opiniões extraídas de um conjunto de *reviews* dado como entrada. Sua estrutura é composta por quatro módulos: **Pré-Processamento** (M1), **Extração de Aspectos** (M2), **Extração de Sentimentos** (M3) e **Sumarização** (M4).

Os módulos são executados de forma independente. As saídas geradas de uns servem de entradas para os outros. Nas seções seguintes, serão descritos cada um dos módulos.

4.1 Módulo de Pré-Processamento

O pré-processamento é uma etapa intermediária que visa preparar os dados para serem utilizados pelo sistema. No caso de informações textuais, essa etapa pode incluir: remoção de texto indesejado, tokenização, remoção de pontuação, remoção de *stopwords* (palavras que não agregam muito significado), uniformização do caso (maiúsculo ou minúsculo), entre outros, dependendo da aplicação.

Especificamente no caso do sistema proposto, o M1 é responsável por adquirir as *reviews* da base de dados, preparar o texto para uso e por fim repassá-lo ao M2. A etapa de aquisição das *reviews* é fundamental. Diferentes bases de dados serão utilizadas, então foi necessário criar um método que unificasse a leitura delas. A solução encontrada foi utilizar a biblioteca ETree da linguagem Python, que permite representar aquivos do formato XML como árvores. As bases de dados utilizadas são todas armazenadas em XML. A aquisição funciona da seguinte maneira: o arquivo XML da base de dados é carregado; em seguida, é necessário especificar em qual *tag* estão as informações requeridas; o M1, utilizando a biblioteca ETree, transforma o arquivo em uma estrutura de árvore e consegue então efetuar a busca baseando-se na *tag* informada. As informações são retornadas para qualquer base de dados organizada em *tags* XML, pois o processo independe do resto da estrutura do arquivo.

Outra funcionalidade de M1 é preparar o texto. Primeiramente, o texto é tokenizado, ou seja, separado em *tokens*. Isso é feito utilizando recursos da linguagem Python. Para o funcionamento do sistema, é necessário remover contrações linguísticas, pontuação, *stopwords* e uniformizá-lo em somente letras minúsculas. Contração, na língua inglesa, é uma versão condensada de palavras ou grupos de palavras (*I've*, por exemplo, é uma contração de *I have*). Ao realizar o pré-processamento, as versões contraídas são substituídas por sua versão longa correspondente. Na versão final do sistema, foram incluídas na lista de contrações algumas abreviações comummente encontradas na Internet – como *bc* para representar *because*. A pontuação é removida utilizando a lista de caracteres que indicam pontuação contida na biblioteca *string*. Cada *token* é comparado à lista e se corresponder a algum elemento, é removido. A remoção de *stopwords* segue o mesmo princípio, mas a lista utilizada é a de *stopwords* da língua inglesa fornecida pela NLTK. A uniformização em letras minúsculas tem como objetivo manter

o texto homogêneo. O funcionamento dessa etapa pode ser resumido pelo Algoritmo 1:

Algoritmo 1: MÉTODO DE PRÉ-PROCESSAMENTO

Entrada: o texto t
Saída: o texto pré-processado t'
início
  $t' = \emptyset$
  Sejam $l1, l2, l3$ as listas de contrações, pontuação e *stopwords*
  para *cada token* $s \in t$ **faça**
    se $s \in l1$ **então**
      $s = s'$
    fim
    se $s \notin l2$ E $s \notin l3$ **então**
      $t' \leftarrow s$
    fim
  fim
fim
retorna t'

O M1 também é responsável por repassar sua saída ao M2. Além do texto pré-processado, o M2 requer mais um tipo de dado para ser executado – a classe gramatical das palavras. O processo de identificar a classe gramatical de uma palavra é conhecido na área de PLN como *part-of-speech tagging* (ou POS *tagging*). Existem diferentes listas que definem quais são as classes gramaticais. As principais são Penn Treebank, que considera 36 classes diferentes, e Brown Corpus, que considera 80, ambas excluindo pontuação. No M1, foram utilizados dois *taggers* diferentes: o fornecido como padrão pela NLTK e o Stanford POS Tagger, ambos baseados nas classes do Penn Treebank. Inicialmente, apenas o da NLTK foi utilizado, mas testes inconsistências nas *tags* obtidas. Foi necessário, então, utilizar o Stanford – que apesar de mais preciso, é muito mais lento. Os testes serão apresentados na Seção 5.

É importante ressaltar que o *tagger* deve ser executado antes da etapa de pré-processamento, pois ele utiliza toda a informação linguística das sentenças para classificar as palavras. Remover *stopwords* ou até mesmo pontuação pode alterar totalmente o resultado. Deve-se, então, primeiramente obter as *tags* e após o pré-processamento remover as que são referentes a palavras que foram excluídas. A saída de M2 é armazenada em arquivo XML contendo a *review* com o texto pré-processado e com as *tags* POS correspondentes. Para exemplificar, considere como entrada a frase "*buy it, love it, and I promise you won't regret it.*". Após o processamento de M1, a saída correspondente é a frase "*buy love promise regret*" junto à lista de *tags* (**NN**, **VBP**, **VBP**, **VB**).

4.2 Módulo de Extração de Aspectos

Esse módulo implementa uma versão adaptada do método proposto por [Hu e Liu 2004], apresentado na Seção 3.3.1.3, o qual apresenta o método de maneira abstrata – permitindo, desse modo, diferentes interpretações computacionais. Sua implementação no sistema apresenta adaptações baseadas em interpretações próprias e no trabalho de [Pavlopoulos e Androutsopoulos 2014]. Todos os algoritmos pertinentes serão descritos em pseudocódigo de modo a facilitar a reprodução do método.

Como entrada, M2 recebe as *reviews* pré-processadas, divididas em sentenças, e as *tags* POS. A leitura do arquivo XML de saída de M1 ocorre de maneira análoga à aquisição da base de dados feita por M1. Depois de lidas, as informações são armazenadas em pares da forma (palavra, *tag*). Uma sentença é uma lista de pares.

Seguindo o método de [Hu e Liu 2004], o primeiro passo é identificar os candidatos a aspectos. Os pares que têm *tag* indicando substantivo são adicionados à lista de candidatos. No Penn Treebank, as *tags* de substantivo são **NN**, **NNS**, **NNP** e **NNPS**. As letras **NN** indicam que a palavra é um substantivo e as demais são detalhes gramaticais não relevantes ao trabalho. A lista de candidatos é expandida combinando os substantivos dois a dois e três a três e os adicionando à lista. A ideia que fundamenta esse processo é que aspectos geralmente são constituídos por substantivos, portanto combiná-los aumenta a chance de encontrar novos aspectos. No entanto, como não há critério para realizar a combinação, muitas vezes o método gera candidatos sem valor semântico e que não pertencem a sentença alguma. Esse procedimento para geração de candidatos advém de [Pavlopoulos e Androutsopoulos 2014], visto que no artigo original os autores não deixam claro qual o método empregado. Apesar de simples, a geração de candidatos pode ser confusa. Para esclarecer o processo, o Algoritmo 2 apresenta detalhadamente todos os

passos:

Algoritmo 2: MÉTODO DE GERAÇÃO DE CANDIDATOS

Entrada: a lista de combinações (palavra, *tag*) m
Saída: a lista de candidatos a aspectos c
início
  $c = \emptyset$
  $subs = \emptyset$
  para *cada elemento* $l \in m$ **faça**
    se $'NN' \in l.tag$ **então**
      $subs \leftarrow l$
    fim
  fim

  para i *de 0* **até** $subs.length$ **faça**
    se $subs[i] \notin c$ **então**
      $c \leftarrow subs[i]$
    fim
    se $(i + 1 < subs.length)\ \&\ (subs[i] + subs[i + 1] \notin c)$ **então**
      $c \leftarrow subs[i] + subs[i + 1]$
    fim
    se $(i + 2 < subs.length)\ \&\ (subs[i] + subs[i + 1] + subs[i + 2] \notin c)$ **então**
      $c \leftarrow subs[i] + subs[i + 1] + subs[i + 2]]$
    fim
  fim
fim
retorna c

O próximo passo é calcular os *p-suportes* de cada termo. Relembrando que o *p-suporte* de um termo t é o número de sentenças que contém t, excluindo as que contém um termo t' que engloba t. Apesar de parecer simples, a implementação do cálculo do *p-suporte* é uma das tarefas mais complexas e importantes do algoritmo. Inicialmente, a solução utilizada era baseada em uma estrutura semelhante a árvores. Porém, ao decorrer do projeto, notou-se que essa estrutura era muito complexa e dificultava o desenvolvimento de determinados métodos. Optou-se, então, por utilizar uma estrutura mais simples que será apresentada posteriormente. Apesar de mais complexa na implementação, a estrutura de árvores facilita a compreensão conceitual do algoritmo e será utilizada para ilustrar o processo. Para exemplificar, considere a frase "*o processador é rápido, mas a duração da bateria é baixa*". Após o pré-processamento e a extração dos substantivos, a frase resultante é "*processador duração bateria*". A partir dessa frase, é possível identificar os candidatos a aspecto *processador*, *duração* e *bateria*; expandindo a lista,

obtém-se: *processador duração*, *duração bateria* e *processador duração bateria*. A estrutura representa os termos candidatos de modo hierárquico de acordo com o número de palavras. Na raiz, fica o candidato com maior número de palavras (que é sempre 3). No nível abaixo, ficam os candidatos de tamanho dois e no último os de tamanho um. Ao avaliar o *p-suporte*, inicia-se pela raiz. Se o termo da raiz for encontrado na sentença, não é necessário avaliar o nível inferior. A mesma ideia se aplica aos demais níveis. A Figura 4.1 ilustra a estrutura para o exemplo anterior.

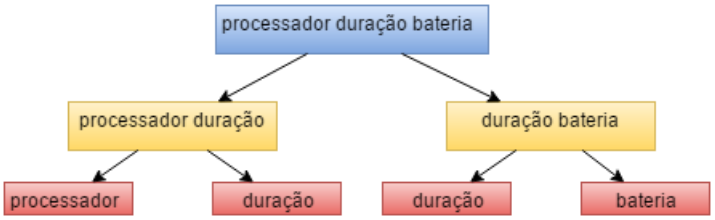

Figura 4.1: Ilustração da estrutura utilizada para calcular o *p-suporte*

Na prática, basta armazenar os termos candidatos em uma lista ordenada pela quantia de palavras que os compõe – ou seja, termos de tamanho 3 são os primeiros. Essa lista será a entrada do algoritmo. Em seguida, deve-se percorrer todas as sentenças para cada termo contido na lista. Se seu tamanho for 3, basta adicioná-lo à lista de *p-suporte* ou incrementar o valor já registrado em seu *p-suporte*, pois termos com esse tamanho não estão contidos em outros. Para termos de tamanho menor, é necessário percorrer a lista de *p-suporte* e verificar se há um termo que o engloba. Caso positivo, o incremento em seu *p-suporte* não é computado. Caso não exista tal termo na lista, o procedimento descrito anteriormente é seguido. O cálculo do *p-suporte* é fundamental para que a extração de aspectos atue corretamente, pois é a base para os demais

passos do método. O Algoritmo 3 descreve em alto-nível a implementação utilizada no sistema:

Algoritmo 3: MÉTODO DE CÁLCULO DO *p-suporte*

Entrada: a lista de termos candidatos ordenada por número de palavras c e a lista de sentenças S
Saída: a lista com o valor de *p-suporte* de cada termo $t \in c$
início
  Seja $p - suporte = \emptyset$ a lista contendo o valor de *p-suporte* de cada termo
  para *cada sentença* $s \in S$ **faça**
    para *cada termo* $t \in c$ **faça**
      se $t \in s$ **então**
        se $t.length == 3$ **então**
          $p - suporte[t] + +$
        fim
        senão
          se $t \notin t'$ ***tal que*** $t.length < t'.length$ **então**
            $p - suporte[t] + +$
          fim
        fim
      fim
    fim
  fim
fim
retorna $p - suporte$

Uma vez calculado o *p-suporte*, é necessário efetuar os métodos de poda – cujo objetivo é reduzir o número de candidatos inválidos. O primeiro é a remoção de termos não-compactos em mais de uma sentença. Para realizar essa verificação, é preciso avaliar as sentenças completas (da saída de M1) que contenham o candidato. Caso a distância entre os termos seja maior que 3 em mais de uma sentença, esse candidato é removido. O segundo método de poda remove termos com *p-suporte* menor que 3 que estão contido em outros. Do ponto de vista prático, termos que não são raiz e têm *p-suporte* menor que 3 são eliminados. Após a poda, todos os candidatos com *p-suporte* superior ao limiar definido são considerados aspectos frequentes. Em geral, o limiar utilizado é 1% do total de sentenças – ou seja, se o candidato está contido em pelo menos em 1% do total de sentenças, ele é considerado um aspecto frequente.

Por fim, os aspectos não-frequentes são extraídos. O cálculo do *p-suporte* possibilita encontrar aspectos que constam em pelo menos uma parte das sentenças. No entanto, há aspectos que são pouco citados nas *reviews* que também podem representar informação de grande valia. Para extrair tais aspectos, não basta diminuir o limiar do *p-suporte* – essa solução aumenta a

quantia de aspectos inválidos e apresenta resultados insatisfatórios. O método empregado no sistema explora a relação entre aspectos e sentimentos. De modo geral, um aspecto (usualmente representado por substantivos) está ligado a um sentimento (representado por adjetivos). Nesse contexto, é possível identificar um aspecto não-frequente avaliando os sentimentos próximos a ele. O algoritmo considera que se há algum adjetivo na sentença com um substantivo vizinho e esse substantivo não consta na lista de aspectos extraídos, ele é então considerado um aspecto não-frequente. Na implementação computacional, primeiramente deve-se identificar todos os adjetivos da sentença por meio das *tags* POS. Na Penn Treebank, adjetivos têm as tags **JJ**, **JJR** e **JJS**. As letra **JJ** indicam que a palavra é um adjetivo e as demais são detalhes gramaticais não relevantes ao trabalho. Em seguida, deve-se localizar todos os substantivos vizinhos aos adjetivos identificados – ou seja, substantivos situados a menos de uma palavra de distância. Por fim, verifica-se se os substantivos localizados já constam na lista de aspectos. Caso negativo, são inseridos e considerados como aspectos não-frequentes. O Algoritmo 4 ilustra esse processo:

Algoritmo 4: MÉTODO DE EXTRAÇÃO DE ASPECTOS NÃO-FREQUENTES

Entrada: a lista de sentenças contendo as *tags* POS m e lista de aspectos frequentes c
Saída: a lista de aspectos não-frequentes f
início
  $f = \emptyset$
  para *cada sentença* $s \in m$ **faça**
    para *cada palavra* $p \in s$ ***tal que*** $'JJ' \in p.tag$ **faça**
      se $\exists n$ ***tal que*** $'NN' \in n.tag \ \ \& \ \ dist(n, p) == 1 \ \ \& \ \ n \notin c$ **então**
        $f \leftarrow n$
      fim
    fim
  fim
fim
retorna f

É importante notar que o algoritmo para extração de aspectos não-frequentes não precisa ser necessariamente aplicado a todas as sentenças da *review*. Em geral, ele é utilizado apenas quando determinada sentença apresenta menos de certo número de aspectos identificados. No sistema, somente sentenças sem aspectos ou com apenas um aspecto frequente passam por esse método. A eficácia do algoritmo é muito dependente da estrutura das *reviews* utilizadas. *Reviews* que contém muitas sentenças sem aspectos podem apresentar resultados ruins, pois o algoritmo tende a gerar muitos falsos positivos – ou seja, considerar como aspectos substantivos

que não deveriam o ser.

A saída de M2 é a lista de sentenças em conjunto com os aspectos identificados para cada uma delas.

4.3 Módulo de Extração de Sentimentos

Esse módulo é responsável por extrair e classificar os sentimentos ligados aos aspectos identificados em M1. Considere que sua entrada é a saída de M2. A abordagem utilizada também é baseada em [Hu e Liu 2004]. O processo é dividido em duas etapas: extração de palavras que indicam sentimento e classificação da polaridade do sentimento.

A primeira segue um procedimento simples: para cada aspecto a obtido de M2, as sentenças são percorridas em busca do adjetivo mais próximo de a. Mais uma vez as *tags* POS são utilizadas. Cada adjetivo encontrado é considerado um sentimento de a. Esse processo é análogo ao descrito no Algoritmo 4.

Na segunda etapa a polaridade dos sentimentos encontrados é definida. O método proposto por [Hu e Liu 2004] utiliza a rede de sinônimos e antônimos da WordNet [University 2015]. Na WordNet, adjetivos são organizados em grupos bipolares. Cada um é conectado a um antônimo, constituindo assim duas polaridades. O adjetivo *rápido*, por exemplo, é conectado a *devagar*, constituindo o grupo bipolar *rápido/devagar*. Além disso, os adjetivos são conectados a sinônimos que expressam o mesmo significado. Em geral, adjetivos têm a mesma polaridade de seus sinônimos e a contrária de seus antônimos. Essa ideia é utilizada para classificar os sentimentos. Inicialmente um conjunto de adjetivos dos quais se sabe a polaridade é usado como base. Ao encontrar um adjetivo não classificado, a WordNet é utilizada e por meio da navegação entre sinônimos e antônimos, o adjetivo tem sua polaridade definida e é adicionado à base. Em [Hu e Liu 2004] 30 adjetivos são adicionados manualmente à base para dar início ao algoritmo.

No sistema implementado, utiliza-se simplesmente uma rede de adjetivos já classificados. Essa rede, criada por [Hu e Liu 2004], conta com mais de 6800 adjetivos com polaridade definida e baseia-se no método descrito acima. Caso algum sentimento encontrado não conste na rede, ele é considerado como neutro.

A saída de M3 é composta pela lista de aspectos obtidos em M2 em conjunto com os sentimentos identificados.

4.4 Módulo de Sumarização

O objetivo desse módulo é gerar resumos contendo as opiniões identificadas para os diferentes aspectos do produto. O módulo visa gerar resumos simples que facilitem a compreensão das informações, sem necessariamente recorrer a algoritmos de sumarização textual. Na implementação atual, são listados os aspectos extraídos (M2), a quantidade de sentimentos positivos e negativos em relação a eles (M3) e um resumo das sentenças completas que os contém, sendo ainda indicado se o aspecto é positivo, neutro ou negativo nessa sentença (M1). Seguindo essa abordagem, é possível filtrar as avaliações de acordo com as características desejadas. Um usuário analisando avaliações sobre um computador pode, por exemplo, aplicar um filtro e ler apenas opiniões positivas a respeito do processador ou apenas negativas a respeito do consumo de energia – ou seja, o usuário não precisa ler todas as avaliações para encontrar a característica que importa para ele.

A estrutura dos sumários gerados pode ser facilmente alterada, visto que a saída dos módulos é armazenada em arquivos .xml. Para gerar sumários diferentes, basta extrair as informações do arquivo e apresentá-las da maneira desejada. Para ilustrar o formato atual, considera-se o aspecto "*duração da bateria*":

Aspecto: **duração da bateria.**

- POS – A **duração da bateria** é **surpreendente**. Também gostei do visor.
- NEUT – Muito bonito esse modelo novo. A **duração da bateria** é **aceitável**, mas o celular esquenta demais
- POS – Muito pesado e caro. A **duração da bateria** é **sensacional**, vale a pena para quem usa o dia inteiro.

Positivo: 2 **Negativo:** 0 **Neutro:** 1

Nota-se que o primeiro atributo mostrado é o aspecto. Em seguida, são apresentadas todas as sentenças que o contém, sendo que no início de cada uma consta a polaridade do aspecto naquela sentença (POS indica positivo, NEUT neutro e NEG negativo). Por fim, apresenta-se a quantidade de sentenças em que o aspecto aparece para cada polaridade.

Capítulo 5

Avaliação Experimental

A avaliação do funcionamento do sistema consiste em comparar os aspectos extraídos automaticamente com aspectos manualmente anotados, ou seja, testar o módulo M2. Serão utilizadas cinco bases de dados com aspectos já anotados. Duas são provenientes de [Pavlopoulos e Androutsopoulos 2014] e três de [Hu e Liu 2004], trabalhos os quais o algoritmo implementado no sistema se baseia. Desse modo, será possível efetuar uma comparação direta de resultados. As bases de dados escolhidas foram:

- **Restaurantes** e **Laptops**: utilizadas em [Pavlopoulos e Androutsopoulos 2014]. Cada uma contém 100 sentenças de *reviews* sobre restaurantes e *laptops*, respectivamente. As *reviews* são curtas e muito ruidosas – com erros ortográficos e palavras irreconhecíveis, principalmente as de *laptops*. Essa inconsistência afeta gravemente o sistema e dificulta a extração de aspectos.

- **Câmera**: contém *reviews* extraídas da Amazon sobre a câmera digital Canon G3, totalizando 530 sentenças. Como as demais abaixo, é utilizada em [Hu e Liu 2004] e apresenta pouco ruído.

- **DVD**: contém *reviews* extraídas da Amazon sobre o aparelho reprodutor de DVD Apex AD2600, totalizando 639 sentenças.

- **Celular**: contém *reviews* extraídas da Amazon sobre o celular Nokia 6610, totalizando 479 sentenças;

No módulo M3, como os sentimentos são classificados com base em uma rede de adjetivos com polaridade já conhecida, não há necessidade em efetuar testes. Durantes alguns testes pre-

liminares dos módulos M1 e M2, ainda na versão parcial do sistema, foi possível perceber que a influência do POS *tagger* utilizado é grande. Como todos os algoritmos implementados são baseados nas classes gramáticas, uma *tag* identificada incorretamente pode prejudicar bastante as demais tarefas. Pôde-se notar que o *tagger* padrão da NLTK cometeu erros em algumas sentenças, principalmente classificando incorretamente substantivos. Por exemplo, na frase "*buy it, love it, and I promise you won't regret it*", "*buy*" foi classificado como substantivo, sendo que na verdade é verbo. O *tagger* de Stanford, apresentou resultados aparentemente superiores, mas seu tempo de execução é elevado. Para auxiliar na decisão sobre qual *tagger* utilizar, foram realizados testes para avaliar a precisão do algoritmo com os diferentes *taggers*. Foram selecionadas para esse teste as bases de dados Restaurantes e Laptops, por apresentarem diferenças mais perceptíveis nas *tags* geradas. A Tabela 5.1 mostra a precisão dos aspectos extraídos com os *taggers* NLTK e Stanford.

Tabela 5.1: Avaliação da Influência dos *Taggers* na Precisão do Algoritmo

Base de Dados	NLTK	Stanford
Restaurantes	55%	56,4%
Laptops	24,3%	33,3%

Apesar de a diferença na base de dados Restaurante ser de apenas 1,4%, na Laptops, que é a mais ruidosa, o *tagger* Stanford apresentou melhora de 9% na precisão. Portanto, optou-se por utilizá-lo nos demais testes.

As métricas escolhidas para avaliar o algoritmo foram *precision* e *recall*. Essas medidas são frequentemente utilizadas para avaliar tarefas de classificação. *Precision* (precisão) indica a fração de instâncias geradas pelo sistema que é considerada relevante – ou seja, a fração do total de aspectos extraídos que realmente é um aspecto válido. *Recall* (revocação) indica a fração de instâncias válidas que foram geradas pelo sistema – ou seja, quantos dos aspectos válidos o sistema foi capaz de extrair. Para calcular essas medidas, os aspectos são divididos em verdadeiro-positivo (vp), falso-positivo (fp) e falso-negativo (fn). Vp indica que o aspecto foi extraído pelo sistema e também consta na base de dados. Fp indica que o aspecto foi extraído, porém, não consta na base de dados, sendo inválido. Fn indica um aspecto que consta na base de dados mas que não foi extraído.

Note que tarefas de classificação usualmente consideram também valores verdadeiro nega-

tivo, porém, essa definição não se aplica ao contexto de extração de aspectos. Logo, o valor de *precision* é dado por $vp/(vp + fp)$ e o de *recall* por $vp/(vp + fn)$, sendo vp o total de aspectos verdadeiro-positivo, fp o de falso-positivo e fn o de falso-negativo.

O algoritmo foi avaliado durante três etapas distintas, com objetivo de investigar o impacto de cada uma no resultado. Foram considerados os aspectos gerados após apenas a etapa de geração de candidatos (doravante denominada E1), após a etapa de poda (E2) e após a identificação de aspectos não-frequentes (E3) – que é a saída final de M2. Em todas as etapas utiliza-se o limiar de 1% para o *p-suporte*. A Tabela 5.2 apresenta o valor de *precision* normalizado (em porcentagem) obtido em cada etapa para cada base de dados.

Tabela 5.2: Porcentagens de *Precision* Resultantes do Algoritmo

Base de Dados	E1	E2	E3
Restaurantes	47,4%	48,2%	56,4%
Laptops	19,1%	21,4%	33,3%
Câmera	51,6%	58,9%	66,7%
DVD	49,1%	57,5%	67,9%
Celular	53,2%	63,1%	68,9%
Média	44,1%	49,82%	58,64%

Observando os resultados é possível notar que a etapa E3 é a mais significativa, acarretando no maior acréscimo de precisão. Os resultados das duas primeiras bases de dados apresentam os piores valores de precisão. Esse fato se deve principalmente ao ruído presente nas sentenças contidas nessas bases de dados. Os erros ortográficos impactam fortemente nos métodos de PLN utilizados no sistema – principalmente no *tagger* POS. Outro fator é que muitas sentenças presentes nessas bases de dados não possuem aspectos. Nesse caso, a etapa de identificação de aspectos não-frequentes vai sempre extrair aspectos que resultam em falso-positivo. A Tabela 5.3 apresenta o valor de *recall* normalizado para as mesmas condições do teste acima.

Tabela 5.3: Porcentagens de *Recall* Resultantes do Algoritmo

Base de Dados	E1	E2	E3
Restaurantes	50,3%	51,4%	64,8%
Laptops	57,4%	59,8%	73,3%
Câmera	62,7%	64,8%	80,3%
DVD	73,1%	74,6%	76,8%
Celular	64,8%	66,4%	75,2%
Média	61,7%	63,4%	74,1%

Analisando os valores de *recall*, confirma-se o impacto da identificação de aspectos não-frequentes em bases de dados com sentenças sem aspectos. Na Laptops, por exemplo, o *recall* foi de 73,3% – ou seja, o algoritmo foi capaz de extrair a maior parte dos aspectos presentes. No entanto, o valor de *precision* foi de 33,3%, menos da metade do *recall*. Isso se deve o fato de o algoritmo gerar muitos candidatos a aspectos não-frequentes e como há sentenças sem aspectos, esses candidatos se tornam falso-positivo e diminuem a precisão da extração. Considerando as *reviews* provenientes da Amazon, os resultados são mais precisos, principalmente pelo texto ser menos ruidoso e conter menos erros de grafia.

Para validar o sistema proposto, os resultados serão comparados aos algoritmos apresentados em [Pavlopoulos e Androutsopoulos 2014] e [Hu e Liu 2004]. A Tabela 5.4 compara os resultados das bases de dados obtidas em [Pavlopoulos e Androutsopoulos 2014]. Apenas o valor de *precision* é avaliado, pois é o único que consta no artigo. A coluna "artigo" representa os valores obtidos pelo algoritmo do artigo e a coluna "proposto" os resultados do algoritmo proposto. A coluna "diferença" contém a diferença entre o resultado do algoritmo proposto e o do artigo. São considerados os resultados finais, ou seja, após todas as etapas de extração de aspectos.

Tabela 5.4: Comparação de Resultados de *Precision* com [Pavlopoulos e Androutsopoulos 2014]

Base de Dados	Artigo	Proposto	Diferença
Restaurantes	52,2%	56,4%	+4,2%
Laptops	34,3%	33,3%	-1%

Pela diferença de resultados, conclui-se que a saída do algoritmo é consistente. Na base de dados Restaurante, o algoritmo apresentou uma precisão maior do que a encontrada no artigo original. Na base de dados Laptops, o resultado obtido pelo algoritmo proposto é 1% menos preciso, porém a diferença é pequena e não compromete a eficácia da extração de aspectos. A versão do algoritmo descrita em [Pavlopoulos e Androutsopoulos 2014] é a que mais se aproxima do algoritmo implementado no sistema, portanto a semelhança nos resultados era esperada. A Tabela 5.5 compara os resultados de *precision* do algoritmo proposto para as bases de dados obtidas em [Hu e Liu 2004] aos resultados do artigo original.

Nesse casos, os resultados obtidos pelo sistema proposto tiveram em média 5,7% a menos de precisão. Como uma parte considerável do algoritmo apresentado em [Hu e Liu 2004] não

Tabela 5.5: Comparação de Resultados de *Precision* com [Hu e Liu 2004]

Base de Dados	Artigo	Proposto	Diferença
Câmera	74,7%	66,7%	-8%
DVD	74,3%	67,9%	-6,4%
Celular	71,8%	68,9%	-2,9%

é explicitada, a interpretação empregada para implementar o sistema proposto pode divergir da original. Outro fator que influencia no resultado é a fase de pré-processamento da base de dados. No artigo original, são utilizados métodos mais complexos para amenizar erros oriundos de problemas presentes na base de dados, como erros ortográficos. Esse tratamento das bases de dados aliado às diferenças de implementação nos algoritmos podem ser a razão da diferença dos resultados. A Tabela 5.6 compara os resultados de *recall* do algoritmo proposto para as bases de dados obtidas em [Hu e Liu 2004] aos resultados do artigo original.

Tabela 5.6: Comparação de Resultados de *Recall* com [Hu e Liu 2004]

Base de Dados	Artigo	Proposto	Diferença
Câmera	82,2%	80,3%	-1,9%
DVD	79,7%	76,8%	-2,9%
Celular	76,1%	75,2%	-0,9%

Para o *recall*, os resultados obtidos pelo sistema proposto apresentaram em média o valor 1,9% abaixo do original. Esse fato mostra que a extração de aspectos contidos nas *reviews* está muito próxima ao algoritmo original, no entanto, a quantidade de aspectos inválidos é maior. Isso se deve em maior parte ao método de identificação de aspectos não-frequentes.

Capítulo 6

Conclusões e Perspectivas

Esse capítulo apresenta uma visão geral sobre o trabalho desenvolvido e discute alguns pontos importantes observados nos experimentos realizados e ao decorrer da revisão bibliográfica. Também são apresentadas propostas de trabalhos futuros e melhorias no sistema proposto.

6.1 Principais Considerações

Na etapa de revisão bibliográfica, foi possível notar que grande parte dos trabalhos publicados concentra-se em discutir técnicas específicas para mineração de opinião, muitas vezes sem contextualizar sua aplicação. Esse trabalho alcançou o objetivo de fornecer um panorama da área, discutindo os temas principais e citando as referências mais relevantes. Apesar de não ser focado diretamente nas técnicas computacionais, esse texto pode servir como uma introdução adequada à área.

Os experimentos realizados com o sistema proposto demonstram que os algoritmos implementados estão consistentes aos já publicados. As diferenças de precisão notadas no processo de extração de aspectos são advindas principalmente do algoritmo de identificação de aspectos não-frequentes. Embora seu objetivo seja gerar candidatos a aspectos visando identificar aspectos efetivamente válidos, muitas vezes o processo acaba gerando candidatos em excesso – reduzindo assim a precisão do sistema. Para amenizar essa adversidade, é necessário refinar o algoritmo. Uma possibilidade é considerar a totalidade de sentenças e avaliar a quantia média de aspectos em cada uma. Com base nesse valor, é possível definir com mais clareza se é necessário efetuar a extração de aspectos não-frequentes. A ideia é que em uma base de dados com muitas sentenças sem aspectos, a média de aspectos por sentença é menor – portanto o

algoritmo não é tão necessário. Outra possibilidade é aplicar métodos de poda aos aspectos não-frequentes, de maneira análoga aos demais aspectos extraídos.

Erros linguísticos presentes nas bases de dados também impactam negativamente na precisão do sistema. Uma abordagem para diminuir o problema é utilizar técnicas de normalização de conteúdo gerado por usuário, que são focadas em melhorar a qualidade de textos extraídos da internet, como em [Clercq et al. 2013].

O sistema proposto também demonstra na prática o processo de mineração de opinião. Todas as etapas, do pré-processamento à sumarização, são explicadas no texto. Desse modo, é possível seguir as etapas descritas sem necessariamente utilizar os mesmos algoritmos. Portanto, esse trabalho também pode ser utilizado como base para implementação de sistemas de mineração de opinião.

6.2 Trabalhos Futuros

Além das sugestões de melhoria do sistema já citadas, algumas possibilidades de trabalhos futuros são:

- Implementar interação com o usuário, permitindo personalização dos sumários gerados.
- Utilizar métodos de normalização textual para melhorar a qualidade das *reviews* utilizadas.
- Implementar um módulo de extração de *reviews*, possibilitando utilizar *reviews* informadas pelo usuário.
- Avaliar e implementar diferentes algoritmos para extração de aspectos, visando melhorar a precisão do sistema.

Referências Bibliográficas

[Amazon 2015]AMAZON. *Apple iPhone 6 Customer Reviews*. 2015. Consultado na Internet: http://www.amazon.com/Apple-iPhone-Silver-16-Unlocked/product-reviews/B00NQGP3L6/ref=cm_cr_dp_qt_see_all_top, 2015.

[Bagheri, Saraee e Jong 2013]BAGHERI, A.; SARAEE, M.; JONG, F. de. An unsupervised aspect detection model for sentiment analysis of reviews. In: MéTAIS, E. et al. (Ed.). *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 7934). p. 140–151. ISBN 978-3-642-38823-1. Disponível em: <http://dx.doi.org/10.1007/978-3-642-38824-8_12>.

[Bakliwal et al. 2013]BAKLIWAL, A. et al. Sentiment analysis of political tweets: Towards an accurate classifier. In: *Proceedings of the Workshop on Language Analysis in Social Media*. Atlanta, Georgia: Association for Computational Linguistics, 2013. p. 49–58. Disponível em: <http://www.aclweb.org/anthology/W13-1106>.

[Berwick 2015]BERWICK, R. *An Idiot's guide to Support vector machines (SVMs)*. 2015. Tutorial. Consultado na Internet: http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf, 2015.

[Bird, Klein e Loper 2009]BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2009. ISBN 0596516495, 9780596516499.

[Blair-Goldensohn et al. 2008]BLAIR-GOLDENSOHN, S. et al. Building a sentiment summarizer for local service reviews. In: *In NLP in the Information Explosion Era*. [S.l.: s.n.], 2008.

[Blei 2015]BLEI, D. M. *Introduction to Probabilistic Topic Models*. 2015. Consultado na Internet: https://www.cs.princeton.edu/ blei/papers/Blei2011.pdf, 2015.

[Blei, Ng e Jordan 2003]BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 993–1022, 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944937>.

[Brooke, Tofiloski e Taboada 2009]BROOKE, J.; TOFILOSKI, M.; TABOADA, M. Cross-linguistic sentiment analysis: From english to spanish. In: *Proceedings of the International Conference RANLP-2009*. Borovets, Bulgaria: Association for Computational Linguistics, 2009. p. 50–54. Disponível em: <http://www.aclweb.org/anthology/R09-1010>.

[Bross 2013]BROSS, J. *Aspect-Oriented Sentiment Analysis of Customer reviews Using Distant Supervision Techniques*. Tese (Doutorado) — Universidade Livre de Berlim, 2013.

[Burges 1998]BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, v. 2, p. 121–167, 1998.

[Clercq et al. 2013]CLERCQ, O. D. et al. Normalization of dutch user-generated content. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, 2013. p. 179–188. Disponível em: <http://www.aclweb.org/anthology/R13-1024>.

[Collins e Singer 1999]COLLINS, M.; SINGER, Y. Unsupervised models for named entity classification. In: *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. [S.l.: s.n.], 1999. p. 100–110.

[Cristianini e Shawe-Taylor 2000]CRISTIANINI, N.; SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. New York, NY, USA: Cambridge University Press, 2000. ISBN 0-521-78019-5.

[Domingos e Pazzani 1997]DOMINGOS, P.; PAZZANI, M. On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 29, n. 2-3, p. 103–130, nov. 1997. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1007413511361>.

[Group 2015]GROUP comScore/Kelsey. *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior*. 2015. Consultado na Internet:

http://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior, 2015.

[Hassan, Qazvinian e Radev 2010]HASSAN, A.; QAZVINIAN, V.; RADEV, D. What's with the attitude?: Identifying sentences with attitude in online discussions. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (EMNLP '10), p. 1245–1255. Disponível em: <http://dl.acm.org/citation.cfm?id=1870658.1870779>.

[Hofmann 1999]HOFMANN, T. Probabilistic latent semantic analysis. In: *Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*. San Francisco, CA: Morgan Kaufmann, 1999. p. 289–296.

[Hogenboom et al. 2013]HOGENBOOM, A. et al. Towards cross-language sentiment analysis through universal star ratings. In: SPRINGER. *7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing*. [S.l.], 2013. p. 69–79.

[Horrigan 2015]HORRIGAN, J. B. *Online Shopping*. 2015. Consultado na Internet: http://www.pewinternet.org/2008/02/13/online-shopping, 2015.

[Hu e Liu 2004]HU, M.; LIU, B. Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2004. (KDD '04), p. 168–177. ISBN 1-58113-888-1. Disponível em: <http://doi.acm.org/10.1145/1014052.1014073>.

[Jiang et al. 2011]JIANG, L. et al. Target-dependent twitter sentiment classification. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (HLT '11), p. 151–160. ISBN 978-1-932432-87-9. Disponível em: <http://dl.acm.org/citation.cfm?id=2002472.2002492>.

[Jin e Ho 2009]JIN, W.; HO, H. H. A novel lexicalized hmm-based learning framework for web opinion mining. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA: ACM, 2009. (ICML '09), p. 465–472. ISBN 978-1-60558-516-1. Disponível em: <http://doi.acm.org/10.1145/1553374.1553435>.

[Jindal e Liu 2006]JINDAL, N.; LIU, B. Mining comparative sentences and relations. In: *Proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence*. Boston, United States: AAAI Press, 2006.

[Jo e Oh 2011]JO, Y.; OH, A. H. Aspect and sentiment unification model for online review analysis. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2011. (WSDM '11), p. 815–824. ISBN 978-1-4503-0493-1. Disponível em: <http://doi.acm.org/10.1145/1935826.1935932>.

[Joachims 1998]JOACHIMS, T. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. 1998.

[Jurafsky e Martin 2009]JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing (2Nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009. ISBN 0131873210.

[Kim e Hovy 2004]KIM, S.-M.; HOVY, E. Determining the sentiment of opinions. In: *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (COLING '04). Disponível em: <http://dx.doi.org/10.3115/1220355.1220555>.

[Lin e He 2009]LIN, C.; HE, Y. Joint sentiment/topic model for sentiment analysis. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2009. (CIKM '09), p. 375–384. ISBN 978-1-60558-512-3. Disponível em: <http://doi.acm.org/10.1145/1645953.1646003>.

[Liu 2012]LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan & Claypool Publishers, 2012.

[Liu, Mobasher e Nasraoui 2011]LIU, B.; MOBASHER, B.; NASRAOUI, O. Web usage mining. In: *Web Data Mining*. Springer Berlin Heidelberg, 2011, (Data-Centric Systems and Applications). p. 527–603. ISBN 978-3-642-19459-7. Disponível em: <http://dx.doi.org/10.1007/978-3-642-19460-3_12>.

[Manevitz e Yousef 2002]MANEVITZ, L. M.; YOUSEF, M. One-class svms for document classification. *J. Mach. Learn. Res.*, JMLR.org, v. 2, p. 139–154, mar. 2002. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944790.944808>.

[McCallum e Nigam 1998]MCCALLUM, A.; NIGAM, K. A comparison of event models for naive bayes text classification. In: *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*. [S.l.]: AAAI Press, 1998. p. 41–48.

[Mooney e Bunescu 2005]MOONEY, R. J.; BUNESCU, R. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 7, n. 1, p. 3–10, jun. 2005. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1089815.1089817>.

[Paetzold 2013]PAETZOLD, G. H. *Um Sistema de Simplificação Automática de Textos escritos em Inglês por meio de Transdução de Árvores*. 2013. Monografia – Universidade Estadual do Oeste do Paraná.

[Pang e Lee 2008]PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, p. 1–135, 2008.

[Pang, Lee e Vaithyanathan 2002]PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (EMNLP '02), p. 79–86. Disponível em: <http://dx.doi.org/10.3115/1118693.1118704>.

[Pavlopoulos e Androutsopoulos 2014]PAVLOPOULOS, J.; ANDROUTSOPOULOS, I. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. In: *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*. [S.l.]: Association for Computational Linguistics, 2014. p. 44–52.

[PHD 2015]PHD, I. *Como funcionam e para que servem as pesquisas de opinião?* 2015. Consultado na Internet: http://www.institutophd.com.br/blog/como-funcionam-e-para-que-servem-as-pesquisas-de-opiniao, 2015.

[Project 2015]PROJECT, N. *Página oficial da plataforma NLTK*. 2015. Consultado na Internet: http://www.nltk.org/, 2015.

[Riloff e Wiebe 2003]RILOFF, E.; WIEBE, J. Learning extraction patterns for subjective expressions. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (EMNLP '03), p. 105–112. Disponível em: <http://dx.doi.org/10.3115/1119355.1119369>.

[Ritter et al. 2011]RITTER, A. et al. Named entity recognition in tweets: An experimental study. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 1524–1534. ISBN 978-1-937284-11-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2145432.2145595>.

[Santos e Ramos 2006]SANTOS, M. Y.; RAMOS, I. *Business Intelligence : tecnologias da informação na gestão de conhecimento*. [S.l.]: FCA - Editora de Informática, 2006.

[Sentiment140 2015]SENTIMENT140. *Sentiment140 website*. 2015. Consultado na Internet: http://http://www.sentiment140.com/, 2015.

[Stamp 2015]STAMP, M. *A Revealing Introduction to Hidden Markov Models*. 2015. Tutorial. Consultado na Internet: https://www.cs.sjsu.edu/ stamp/RUA/HMM.pdf, 2015.

[Stats 2015]STATS, I. L. *Twitter live stats*. 2015. Consultado na Internet: http://www.internetlivestats.com/twitter-statistics, 2015.

[Turney 2002]TURNEY, P. D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ACL '02), p. 417–424. Disponível em: <http://dx.doi.org/10.3115/1073083.1073153>.

[Tweetfeel 2014]TWEETFEEL. *Tweetfeel website*. 2014. Consultado na Internet: http://www.crunchbase.com/organization/tweetfeel, 2014.

[Twtbase 2015]TWTBASE. *Twitrratr website*. 2015. Consultado na Internet: http://www.twtbase.com/twitrratr, 2015.

[University 2015]UNIVERSITY, P. *WordNet: A lexical database for English*. 2015. Consultado na Internet: https://wordnet.princeton.edu/, 2015.

[Walmart 2015]WALMART. *Avaliações de consumidor sobre o produto Tablet Samsung Galaxy Tab*. 2015. Consultado na Internet: https://www.walmart.com.br/tablet-samsung-galaxy-tab-s-sm-t700n-tela-8-4-android-4-4-16gb-wi-fi-branco-octa-core-de-1-9ghz-1-3ghz/2448436/pr?pageNumber=1, 2015.

[Wang e Ren 2015]WANG, J.; REN, H. *Feature-Based Customer Review Mining*. 2015. Consultado na Internet: http://nlp.stanford.edu/courses/cs224n/2007/fp/johnnyw-hengren.pdf, 2015.

[Wiebe 2000]WIEBE, J. Conference review. *Intelligence*, ACM, New York, NY, USA, v. 11, n. 2, p. 43–48, jun. 2000. ISSN 1523-8822. Disponível em: <http://doi.acm.org/10.1145/337897.338001>.

[Wiebe e Riloff 2005]WIEBE, J.; RILOFF, E. Creating subjective and objective sentence classifiers from unannotated texts. In: GELBUKH, A. (Ed.). *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2005, (Lecture Notes in Computer Science, v. 3406). p. 486–497. ISBN 978-3-540-24523-0. Disponível em: <http://dx.doi.org/10.1007/978-3-540-30586-6_53>.

[Wiebe e Riloff 2005]WIEBE, J.; RILOFF, E. Creating subjective and objective sentence classifiers from unannotated texts. In: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer-Verlag, 2005. (CICLing'05), p. 486–497. ISBN 3-540-24523-5, 978-3-540-24523-0. Disponível em: <http://dx.doi.org/10.1007/978-3-540-30586-6_53>.

[Wiebe, Bruce e O'Hara 1999]WIEBE, J. M.; BRUCE, R. F.; O'HARA, T. P. Development and use of a gold-standard data set for subjectivity classifications. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*.

Stroudsburg, PA, USA: Association for Computational Linguistics, 1999. (ACL '99), p. 246–253. ISBN 1-55860-609-3. Disponível em: <http://dx.doi.org/10.3115/1034678.1034721>.

[Yu e Hatzivassiloglou 2003]YU, H.; 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*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (EMNLP '03), p. 129–136. Disponível em: <http://dx.doi.org/10.3115/1119355.1119372>.

[Zaki e Jr 2014]ZAKI, M. J.; JR, W. M. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York, NY, USA: Cambridge University Press, 2014. ISBN 0521766338, 9780521766333.

[Zhang e Liu 2011]ZHANG, L.; LIU, B. Identifying noun product features that imply opinions. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistic*. Oregon, United States: Association for Computational Linguistics, 2011.

[Zoghbi, Vulić e Moens 2013]ZOGHBI, S.; VULIć, I.; MOENS, M.-F. I pinned it. where can i buy one like it?: Automatically linking pinterest pins to online webshops. In: *Proceedings of the 2013 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media*. New York, NY, USA: ACM, 2013. (DUBMOD '13), p. 9–12. ISBN 978-1-4503-2417-5.