Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Um estudo experimental para extração de características de textos opinativos em língua portuguesa.

Bruno Pelizari Dutra Pettersen

JUIZ DE FORA
DEZEMBRO, 2014

Um estudo experimental para extração de características de textos opinativos em língua portuguesa.

Bruno Pelizari Dutra Pettersen

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Jairo Francisco de Souza

JUIZ DE FORA

DEZEMBRO, 2014

UM ESTUDO EXPERIMENTAL PARA EXTRAÇÃO DE CARACTERÍSTICAS DE TEXTOS OPINATIVOS EM LÍNGUA PORTUGUESA.

Bruno Pelizari Dutra Pettersen

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Jairo Francisco de Souza
Doutor em Informática/ PUC-Rio

Luciana Brugiolo Gonçalves
Doutora em Ciência da Computação/UFF

Luciana Conceição Dias Campos
Doutora em Engenharia Elétrica/PUC-Rio

JUIZ DE FORA

18 DE DEZEMBRO, 2014

Resumo

A área de análise de sentimentos visa coletar, extrair, classificar e sumarizar opiniões de pessoas em relação a alguma entidade específica. Para que o processo seja válido e significativo, cada etapa da análise deve ser realizada com o máximo rigor. Devido à língua portuguesa ser dinâmica, há a possibilidade de se construir uma frase de mesmo significado semântico de diversas maneiras, o que torna a etapa de extração de características um desafio. Este trabalho apresenta um estudo experimental para extração de características de textos opinativos em língua portuguesa. No estudo apresentado, uma abordagem foi desenvolvida com base em propostas presentes na literatura e adaptados para o contexto do trabalho. O projeto tem como objetivo melhorar a identificação de características de textos opinativos em língua portuguesa, possibilitando uma classificação mais precisa para as fases seguintes do processo de análise de sentimentos. A abordagem foi avaliada através da comparação de um *benchmark* com textos opinativos sobre produtos e serviços.

Palavras-chave: Análise de Sentimentos, Extração de Características , Processamento de Linguagem Natural.

Abstract

The sentiment analysis area aims to collect, extract, classify and summarize opinions of people in relation to any particular entity. For the process to be valid and meaningful, every step of the analysis should be performed with the utmost rigor. Because the portuguese language is dynamic, there is the possibility of constructing a sentence of the same semantic meaning in different ways, which makes the feature extraction step a challenge. This paper presents an experimental study to feature extraction of reviews written in portuguese. In the study, an approach was developed based on proposals in the literature and adapted to the work context. The project aims to improve the identification of features in opinion pieces in portuguese, enabling a more precise classification for the following phases of sentiment analysis process. The approach was evaluated by comparing a benchmark with opinion pieces about products and services.

Keywords: Sentiment Analysis, Feature Extraction, Natural Language Processing.

Agradecimentos

Agradeço todo o apoio dos meus pais e meu irmão, que me deram força e incentivo ao longo do curso inteiro. E que foram muito compreensíveis com a minha ausência em vários momentos. Amo vocês.

A todos parentes e amigos, por toda torcida pelo meu êxito.

Ao professor Jairo Francisco pela orientação, amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria. Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

Gostaria também de agradecer a importante contribuição do bolsista Thiago Goldoni, que foi muito generoso ao me ajudar na etapa de testes do protótipo.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Com o avanço tecnológico dos últimos tempos, computadores ficaram com maior capacidade de processamento e armazenamento. As redes de telecomunicações melhoraram suas infraestruturas e o setor de dispositivos móveis com acesso a *Internet* cresceu de forma vertiginosa. Como resultado, as pessoas passaram a estar cada vez mais conectadas à rede, não mais como observadoras e consumidoras de conteúdo, como ocorreu no início da *Internet*, mas também como fornecedoras do mesmo, através de comentários em redes sociais, *blogs*, fóruns, etc.

De acordo com Vilicic (2013), estima-se que diariamente sejam produzidos aproximadamente 2,5 exabytes[1] de informações pela humanidade. Fazendo uma comparação, a cada quinze minutos, a humanidade gera o triplo de informações disponíveis no acervo da biblioteca do congresso americano, a maior do mundo. Eis o cenário do *"Big Data"*, que se caracteriza pela abundância de dados, sua variedade e a velocidade com que trafegam no universo digital.

Essa abundância de dados pode parecer aparentemente sem valor para um leigo, só que traz uma rica quantidade de informações, que são úteis tanto para empresas como consumidores. Por exemplo, uma empresa pode através da análise de tais dados, ter uma visão geral sobre a imagem de sua marca, pode acompanhar a aceitação de um produto lançado no mercado, como consumidores também podem utilizar os dados para decidir se devem comprar um produto x ou um produto y.

Acontece que, à medida que esses dados crescem exponencialmente, fica praticamente impossível extrair informações de forma manual. Talvez um consumidor lendo algumas opiniões na internet, possa definir qual seja a melhor decisão de compra para ele. O mesmo já não é válido para uma empresa. Coletar tais informações de forma manual e abrangente, para ter uma visão de seu público consumidor, seria uma tarefa inviável. Na verdade, é necessário que esse processo seja automatizado através de algoritmos (Liu, 2012).

[1] 1 exabyte corresponde a 1,15 x $10^{18} bytes$.

A área de Análise de Sentimentos surgiu para suprir essa necessidade. É uma área que passou a ser estudada recentemente e que busca aprimorar técnicas para coletar, extrair, classificar e sumarizar a opinião das pessoas em relação a uma determinada entidade a partir de textos de forma automatizada.

Neste projeto o foco será dado a uma etapa da Análise de Sentimentos em especial, a etapa de extração de características. Essa etapa é fundamental para o processo de Análise de Sentimentos, é nela que se identifica e extrai-se o alvo de uma opinião. Logo, se realizada de forma inadequada, pode comprometer todo o processo de análise. Assim, propõem-se desenvolver um protótipo que utilize abordagens já propostas na literatura buscando encontrar novas técnicas que promovam melhorias nessa etapa.

1.1 Justificativa

De acordo com Liu (2012) as opiniões são fundamentais para quase todas as atividades humanas, porque elas são os principais influenciadores do comportamento humano. Antes de a *Internet* evoluir para o cenário que temos atualmente, as pessoas eram influenciadas geralmente pela opinião de membros próximos ligados a elas, tais como amigos e parentes. Com o avanço da *Internet*, as pessoas estão cada vez mais conectadas a rede. Pode-se dizer que agora, além de membros conhecidos, as pessoas também ouvem a opinião de pessoas totalmente desconhecidas.

A *Internet* tornou-se um imenso repositório de dados. Indivíduos através de redes sociais, foruns, *blogs*, fazem críticas sobre pessoas, produtos, organizações. Comentam fatos relacionados ao seu espaço geográfico, tais como: ocorrência de catástofres, terremotos, congestionamentos. Expõem desejos pessoais sobre mudança em determinados produtos ou serviços, revelam dados pessoais, etc. Enfim, a *Internet* tornou-se um veículo de comunicação e relacionamento onde as pessoas sentem-se livres para compartilhar informações, de uma forma espontânea sem ter que se preocuparem tanto em serem identifidadas.

E esses dados na rede são extremamente valiosos. Ao fazer uma compra, escolher um destino de viagem, verificar a idoneidade de uma empresa, pessoas cada vez mais buscam informações na rede e dados positivos e negativos são levados em consideração e são determinantes em uma decisão final. Logo, essas informações são também de grande

interesse das empresas. Uma imagem negativa de uma empresa ou produto na *Internet* é algo muito prejudicial. Ninguém ao fazer uma pesquisa sobre um produto e descobrir que ele é alvo de muitas críticas irá ter a mesma propensão a comprá-lo.

Observando sob este prisma, as empresas estão cada vez mais interessadas em saber o que as pessoas estão pensando em relação a elas a partir dos dados publicados na rede, desejam acompanhar a aceitação no mercado de um produto lançado, críticas em relação aos serviços prestados, identificar concorrentes, captar os desejos dos consumidores, enfim, querem extrair o máximo de informações possíveis para que ao serem analisadas por um especialista no assunto, possam ser uma base para a tomada de decisões estratégicas.

Dada essa entre outras necessidades, a Mineração de Textos passou, portanto, a ser foco de diversas pesquisas. Existem atualmente algumas tarefas importantes que compõem a área. Uma delas que vêm sendo bastante explorada é a Análise de Sentimentos, que busca coletar mensagens opinativas de diversas origens na *Internet* em relação a determinado alvo, avaliar a polaridade dos sentimentos expressos nas mesmas e por fim sumarizar os resultados obtidos.

Como a linguagem natural é composta por um conjunto muito grande de regras gramaticais, que variam conforme o idioma, e dado o seu caráter flexível, onde o escritor pode expressar suas ideias em uma sentença de diversas maneiras, a tarefa de minerar textos não representa algo trivial. De acordo com Liu (2012), vários trabalhos têm sido desenvolvidos em relação à Área de Análise de sentimentos Baseada em Aspectos, mas dada a complexidade da tarefa, a maioria deles ainda apresentam resultados pouco satisfatórios.

Tendo como base esse cenário onde as pesquisas ainda não atingiram grande evolução, nesse projeto busca-se fazer um estudo da área de Análise de Sentimentos Baseada em Aspectos em busca de melhorias, sendo estas voltadas especificamente para a etapa de extração de características. E como é uma área relativamente nova, existem poucos trabalhos relacionados, principalmente para língua portuguesa. Logo, foi o idioma adotado nesta pesquisa.

Como no processo de análise de sentimentos as etapas geralmente são sequenciais, é fundamental que a etapa de extração de características seja bem realizada para que os

resultados finais possam ser confiáveis e não haver perdas de informação.

A tendência é que, com o crescimento do comércio eletrônico, cada vez mais as técnicas de análise de sentimentos sirvam de suporte tanto para as empresas no monitoramento de suas marcas, quanto para os clientes que utilizarão as informações sumarizadas para decidirem quanto à aquisição de produtos e serviços. Assim, pode-se dizer que é uma área de pesquisa importante e que trará muitas vantagens competitivas para as empresas que possuirem tais sistemas.

1.2 Objetivos

O objetivo principal deste projeto é propor melhorarias à extração de características em textos opinativos em língua portuguesa. Para tal será realizado um estudo experimental.

Como objetivo específico, será criado um protótipo com o qual se realizará todos os testes para validação das abordagens adotadas.

Além disso, o trabalho objetiva criar uma base de avaliações de textos opinativos para a língua portuguesa, facilitando a comparação entre diferentes abordagens futuras.

Espera-se que o protótipo final seja capaz de cumprir sua tarefa a partir de textos opinativos de diversos domínios e que apresente uma boa acurácia nos testes realizados.

1.3 Estrutura do Trabalho

O trabalho apresenta cinco capítulos, descritos a seguir:

O Capítulo 1 representa a parte introdutória do trabalho onde é exposta a área de pesquisa, a proposta do projeto juntamente com os objetivos a serem alcançados e a razão pela qual o tema merece ser abordado.

No Capítulo 2 é definido o estado de arte do problema abordado, sendo citados os principais trabalhos relacionados desenvolvidos e suas respectivas técnicas. Assim como os principais conceitos da área.

O Capítulo 3 descreve o desenvolvimento do protótipo, apresentando os módulos que o compõem e detalha as técnicas e procedimentos adotados.

O Capítulo 4 representa a parte de validação do protótipo desenvolvido, onde são descritos os testes efetuados com o objetivo de avaliar a eficiência do mesmo na tarefa de extrair características e suas respectivas expressões opinativas. Os resultados finais são apresentados e comentados.

No Capítulo 5 são feitas as considerações finais, apresentadas as contribuições para a área de Análise de Sentimentos, bem como as dificuldades encontradas para realização do trabalho. E por fim é apresentada uma relação de possíveis melhorias em pesquisas futuras.

2 Análise de Sentimentos

A seguir são definidos os principais conceitos da área, nomenclaturas utilizadas neste trabalho, o estado de arte da Análise de Sentimentos e alguns trabalhos relacionados a este projeto.

2.1 Conceitos Básicos

A análise de sentimentos é um campo de pesquisa relativamente novo que se iniciou por volta dos anos 2000 e que foi profundamente impulsionado pelo grande volume de opiniões geradas por redes sociais, *blogs*, fóruns, etc. Aliado, principalmente, ao grande interesse das empresas em extrair informações estratégicas desse vasto conjunto de dados.

Conceitualmente, a análise de sentimentos pode ser realizada em três níveis: documento, sentença ou por entidade e aspecto (Liu, 2012). Abaixo é apresentado o grau de detalhamento de cada nível:

- no **nível de documento**, a tarefa é classificar a opinião geral do mesmo como um sentimento positivo ou negativo (Turney, 2002). Esse tipo de análise é realizado somente em textos que expressam opiniões sobre apenas uma entidade, não sendo possível analisar textos que fazem a comparação de várias entidades;

- no **nível de sentença**, a tarefa é determinar a cada sentença se o sentimento expresso é positivo, negativo ou neutro. Neutro significa sem opinião. Este nível de análise está intimamente relacionado à classificação de subjetividade (Wiebe et al, 1999), que diferenciam sentenças objetivas, que expressam informações concretas, de sentenças subjetivas, que expressam visões pessoais;

- por fim tem-se o **nível de entidade e aspecto**. Este nível apresenta uma maior granularidade. A análise parte do pressuposto que uma opinião consiste de sentimentos positivos e negativos relacionados a determinado alvo, que representa o objeto da opinião. Assim, pode-se descobrir o que exatamente pessoas aprovam ou

desaprovam, diferentemente das análises em nível de documento e de sentença que se limitam a analisar a positividade ou negatividade de um texto ou sentença, porém não relacionam esses sentimentos a um alvo específico.

Existem algumas palavras que tipicamente expressam sentimentos positivos ou negativos, tais como adjetivos e advérbios. A análise dessas palavras é importante no processo de análise de sentimentos, porém não é suficiente para a resolução do problema em questão, que é muito mais complexo. Algumas questões envolvidas são:

- **palavras de sentimento podem ter orientações opostas de acordo com o domínio da aplicação.** Conforme exemplificado abaixo:

"A pizza estava gelada." A palavra opinativa "gelada" no domínio de "pizza" tem uma polaridade negativa. "A cerveja estava gelada." A palavra opinativa "gelada" no domínio "cerveja" possui uma polaridade positiva.

- **sentenças contendo palavras que geralmente expressam sentimentos podem na verdade não expressar qualquer sentimento.** Geralmente ocorre em sentenças condicionais e interrogativas. Conforme exemplificado abaixo:

"Se encontrar um bom preço de venda do iPhone 5S eu comprarei." "O iPhone 5S é um bom celular ?"

  Nas duas sentenças existe a presença da palavra de sentimento "bom", porém ela não expressa sentimentos positivos ou negativos relativos ao iPhone 5S. Mas existem sentenças condicionais e interrogativas que expressam sentimentos. Conforme exemplificado abaixo:

"Se você quiser um bom celular, compre o iPhone 5S." "Como faço para enviar fotos nesse terrível celular Xing Ling?"

- o **emprego de ironias** também dificulta a análise, pois são empregadas palavras de sentimento que geralmente representam a polaridade oposta do que a que o escritor realmente gostaria de remeter. Conforme exemplificado abaixo:

"Que ótimo celular, muito prático!!!! Tenho que carregar sua bateria 3 vezes por dia."

- **sentenças que não possuem palavras de sentimento podem representar opiniões.** Geralmente representam sentenças objetivas, que apresentam apenas fatos ocorridos. Conforme exemplificado abaixo:

"O celular não liga após uma semana de uso."

"Esse carro consome muito combustível."

Após terem sido abordados os níveis do processo Análise de sentimentos e algumas questões que dificultam o processo, alguns conceitos são importantes serem introduzidos. O primeiro deles refere-se à definição de opinião. Segundo o Dicionário do Aurélio (2014) opinião é: "Maneira de pensar particular sobre um assunto". "Julgamento, conceito favorável ou não sobre uma pessoa, um assunto, uma coisa".

De acordo com Jindal et al (2006) há dois principais tipos de opiniões: regulares e comparativas. Opiniões regulares são expressões de sentimento sobre alguma entidade alvo. Exemplo:

"Essa televisão possui uma qualidade de imagem fantástica."

Já as opiniões comparativas ocorrem em expressões que referenciam mais de uma entidade. Exemplo:

"O design dos carros da Peugeot é melhor que os da Fiat."

Entidades e aspectos são outros conceitos importantes. Segundo Hu et al (2004) entidade e é um produto, pessoa, evento, organização ou tópico. Sendo e representado como:

- uma hierarquia de componentes, subcomponentes e assim por diante;
- cada nó representa um componente e está associado a um conjunto de atributos do componente;

- aspectos (características) representam ambos, componentes e atributos. São considerados alvos de uma opinião, ou seja, sobre o que uma expressão opinativa modifica;
- Uma opinião pode ser expressa sobre qualquer nó ou atributo do nó.

Para evitar qualquer confusão, a nomenclatura "características"e "aspectos"utilizadas neste projeto, representam sinônimos e são intercambiáveis.

Existem dois tipos de características:

- **características explícitas:** Representados por substantivos e locuções substantivas.

  "A qualidade de imagem da câmera X é ótima."

  Neste caso "qualidade de imagem"representa uma característica explícita. A opinião nesta sentença está sendo feita portanto, em relação a qualidade da imagem (que representa um característica) da câmera X (que representa uma entidade). A opinião atribuída é positiva, sendo feita através da palavra opinativa "ótima";

- **características implícitas:** Representadas mais frequentemente por adjetivos e advérbios, mas que podem ser representadas também por verbos. Essas palavras quando encontradas em um texto podem referenciar uma característica, que neste caso está implícita.

  No exemplo abaixo, o adjetivo "caro"referencia uma característica implícita, neste caso, "preço".

  "Este produto é muito *caro*."

  Não é a característica "produto"que é "caro"e sim a característica "preço". A identificação de tais características é importante no processo de análise de sentimentos, pois promove uma maior granularidade na análise.

  Se for considerado que "produto"é "caro", a atribuição não deixa de fazer sentido, porém a contagem de sentimentos, positivo neste caso, irá ser dada a "produto"e tem-se a noção de que "produto"é uma característica positiva como um

todo. Quando na verdade a opinião é mais restrita e refere-se somente a um atributo da entidade "produto", logo o "preço"do produto. Que poderia ser: a dimensão do produto, a resistência do produto, etc. Logo, identificando tais características melhora-se o nível de detalhamento da análise.

Uma formalização da definição de opinião é proposta por Liu (2010). Uma opinião é representada por uma quíntupla: *(e, a, s, h, t)*, onde:

- e: representa uma entidade;
- a: representa um aspecto;
- h: representa um opinante;
- t: representa o tempo em que a opinião foi expressa;
- s: representa o valor da opinião, feita pelo opinante h, sobre o aspecto a, pertencente a entidade e, no tempo t. Sendo, positivo, negativo, neutro, ou recebe uma classificação mais granular.

Definidos os conceitos principais, pode-se dizer que o objetivo da análise de sentimentos é: dado um documento, descobrir todas as quíntuplas (opiniões) contidas no mesmo. A seguir um exemplo que representa um comentário de um produto:

Postado por: Bruno Pelizari Data:20/04/2014 (1) Semana passada eu comprei um Peugeot 308. (2) Realmente o 308 é um excelente carro. (3) Possui um motor potente e apresenta boa estabilidade nas curvas. (4) Além disso, é muito confortável. (5) Porém achei que o design da parte traseira poderia ser melhor. (6) Minha namorada achou o preço um pouco elevado.

Têm-se os seguintes passos para extração das quíntuplas:

- **extração de entidades e categorização:** Extraem-se todas as entidades contidas em cada documento e agrupam-se as entidades que representam sinônimos no mesmo *cluster* ou mesma categoria. Assim cada *cluster* representa uma entidade única. Exemplo:

“Pegeout 308” e “308” são agrupadas no mesmo *cluster*, pois, representam a mesma entidade.

- **extração de aspectos e categorização:** Extraem-se todos os aspectos das entidades e agrupam-se os mesmos em *clusters*. Cada *cluster* de uma entidade representa um aspecto único. Exemplo:

Sentenças (3) “motor”, “estabilidade nas curvas” (4)"conforto"(5) “design da parte traseira” (6) “preço”

- **extração do detentor da opinião e categorização.** Saída abaixo:

Sentenças (2), (3), (4), (5) tem como opinante Bruno Pelizari. Sentença (6) tem como opinante a namorada de Bruno Pelizari

- **extração da data da opinião e padronização para um formato definido de data.** Saída abaixo:

A data do comentário foi: 20/04/2014

- **classificação do aspecto:** Determinar se uma opinião sobre um determinado aspecto é positiva, negativa ou neutra, ou atribuir um valor numérico para o aspecto. Saída abaixo:

Na sentença (2) há uma opinião positiva em relação à entidade “Peugeot 308”
Na sentença (3) há uma opinião positiva em relação aos aspectos “motor” e “estabilidade na curva”
Na sentença (4) há uma opinião positiva em relação ao aspecto implícito “conforto”
Na sentença (5) há uma opinião negativa em relação ao aspecto “design da parte traseira”
Na sentença (6) há uma opinião negativa em relação ao aspecto “preço”

- **geração das quíntuplas:** Extração de todas as quíntuplas de determinado documento através dos resultados dos passos anteriores.

  Na primeira quíntupla gerada o aspecto é "geral", essa designação é dada quando o alvo de uma opinião é uma entidade.

(Peugeot 308, geral, positivo, Bruno Pelizari, 20/04/2014)
(Peugeot 308, motor, positivo, Bruno Pelizari, 20/04/2014)
(Peugeot 308, estabilidade nas curvas, positivo, Bruno Pelizari, 20/04/2014)
(Peugeot 308, conforto, positivo, Bruno Pelizari, 20/04/2014)
(Peugeot 308, design da parte traseira, negativo, Bruno Pelizari, 20/04/2014)
(Peugeot 308, preço, negativo, namorada do Bruno Pelizari, 20/04/2014)

Na análise de sentimentos a opinião de apenas uma pessoa geralmente não é relevante (a menos que ela exerça grande influência – *VIP* [2]. Portanto é necessário que se reúna a opinião de várias pessoas e que estas sejam sintetizadas. Com o uso das quíntuplas definidas anteriormente pode-se gerar um resumo tanto qualitativo quanto quantitativo das informações obtidas e é através destes dados finais que o gestor de uma empresa (especialista no negócio) irá se basear de forma a auxiliá-lo na tomada de decisões.

Após terem sido introduzidos os conceitos iniciais sobre a Análise de Sentimentos, nas próximas seções será abordado o estado de arte da Extração de Características e Palavras Opinativas, que compõe uma das etapas da Análise de Sentimentos Baseada em Aspectos e é o foco deste trabalho.

Como visto anteriormante, existem dois tipos de características: as explícitas e as implícitas. Primeiro será descrito as abordagens utilizadas para extração das características explícitas.

2.2 Extração das Características Explícitas

Para essa extração existem quatro abordagens principais: extração baseada na frequência de substantivos e locuções substantivas, extração pela exploração da opinião e seus relacionamentos, extração usando aprendizado supervisionado e extração usando modelagem de tópico. A seguir são descritas cada abordagem:

[2] *Very Important Person*: refere-se a pessoas muito influentes.

2.2.1 Extração baseada na frequência de substantivos e locuções substantivas

A abordagem descrita por Hu et al (2004), baseia-se na frequência de substantivos e locuções substantivas. Leva em consideração que quanto mais frequentes estes termos forem encontrados em um documento, maior será a chance dos mesmos representarem aspectos relacionados à determinada entidade. A identificação desses elementos gramaticais foi realizada por um *POS Tagger*[3], que classificou gramaticalmente todo texto de entrada e em seguida computou-se a frequência de cada substantivo e locução substantiva. As maiores frequências, dentro de um limiar que foi definido experimentalmente, representaram os prováveis candidatos a aspectos do corpus. Os que apresentaram frequências mais baixas, não foram considerados aspectos ou foram considerados de menor importância. É considerado um método simples e bastante efetivo.

Alguns trabalhos realizaram melhoramentos nessa técnica e obtiveram resultados mais precisos. Um exemplo foi o trabalho de Popescu et al (2005), a qual se utilizou do cálculo do *PMI*[4] para selecionar entre os candidatos a aspectos determinados pela frequência dos termos, quais possivelmente foram classificados erroneamente, não sendo aspectos verdadeiros de determinada entidade.

2.2.2 Extração pela exploração da opinião e seus relacionamentos

A segunda abordagem, usando opinião e alvo relacionado, foi utilizada no trabalho de Hu et al (2004) para extrair aspectos infrequentes. A ideia central do trabalho é que uma mesma palavra de sentimento pode ser usada para descrever ou modificar diferentes aspectos. Logo se uma sentença não possui um aspecto frequente, mas possui alguma palavra de sentimento, o substantivo ou locução substantiva mais próxima de cada palavra de sentimento é extraído e classificado como um candidato a aspecto.

Existe uma técnica disponível na literatura que utiliza as duas primeiras abordagens, chamada *Double Propagation*, que realiza simultaneamente a extração de aspectos

[3]Part-of-speech Tagger : ferramenta usada para assinalar as categorias gramaticais a cada palavra em um *corpus*.

[4]*Pointwise Mutual Information*: medida de similaridade utilizada para determinar o grau de associação entre palavras.

e palavras de sentimento. Considerando que cada opinião sempre possui um alvo, (Qiu et al, 2009), obteve tal resultado explorando certas relações sintáticas entre sentimentos e alvos. Seu algoritmo teve como entrada um pequeno grupo de palavras de sentimento sementes, e não foram requeridos aspectos como sementes.

A ideia desenvolvida foi a seguinte, como palavras de sentimento e alvos possuem um relacionamento entre si, pode-se a partir de aspectos identificados reconhecerem palavras de sentimento. O contrário também é verdadeiro, a partir de palavras de sentimento conhecidas podem-se identificar os aspectos relacionados. E através de um processo iterativo, são determinados a cada palavra de sentimento e aspecto extraído, novos aspectos e palavras de sentimento até que a condição de parada do algoritmo retorne vazio. A extração segue regras de dependências gramaticais e o método tem como vantagem a independência de domínio.

2.2.3 Extração usando aprendizado supervisionado

A terceira abordagem, a partir de aprendizado supervisionado, utiliza majoritariamente métodos baseados em aprendizado sequencial, tais como *HMM (Hidden Markov Models)* (Rabiner, 1989) e *CRF (Conditional Random Fields)* (Lafferty et al, 2001). Essas técnicas necessitam que aspectos e não aspectos sejam identificados manualmente no corpus para serem usados para treinamento, o que pode ser bastante custoso.

No trabalho de Jakob et al (2010), utilizou-se o *Conditional Random Fields*. O modelo foi treinado a partir de sentenças de diversos domínios em busca de um extrator de características mais robusto, minimizando a dependência de domínio. Um grupo de características independentes do domínio foi utilizado: tokens, dependência sintática, distância das palavras, *POS Taggers* e sentenças de opinião.

Outro trabalho desenvolvido por Kobayashi et al (2007), numa primeira etapa encontra-se pares de candidatos a aspectos e a opinião correspondente usando uma árvore de dependência. Em seguida emprega-se um método de classificação por árvore estruturada para aprender e classificar os pares candidatos, que apresentaram uma maior pontuação, como sendo os aspectos. As características consideradas no aprendizado foram: dicas contextuais, dicas estatísticas de co-ocorrência, entre outras.

2.2.4 Extração usando modelagem de tópico

A quarta abordagem, usando modelagem de tópicos, é um método de aprendizado não supervisionado onde cada tópico contido em um documento representa uma distribuição de probabilidade sobre as palavras. O algoritmo atua formando *clusters*, sendo cada *cluster* formado por um conjunto de palavras que representam determinado tópico. Onde tópico, neste contexto, é sinônimo de aspecto. Existem dois modelos principais: *pLSA (Probabilistc Latente Semantic Analysis)* (Hofmann, 1999) e *LDA (Latent Dirichlet Allocation)* (Blei et al, 2003).

O método básico extrai ao mesmo tempo aspectos e palavras de sentimento. Para a análise de sentimentos esses dois elementos precisam estar separados. Assim pode-se estender o método básico de tal forma a adaptá-lo, extraindo os dois elementos, mas distinguindo ambos.

Como ponto negativo dessa abordagem, pode-se levar em conta que a mesma requer um grande volume de dados e de significativos ajustes para que obtenha resultados satisfatórios. Apresenta grande capacidade para extrair aspectos gerais e frequentes em um documento, mas tem limitações para extrair aspectos que são frequentes apenas localmente, os quais muitas vezes são os mais relevantes para entidades específicas que o usuário pode estar interessado.

De uma forma geral esta abordagem não é suficientemente granular e específica para a maioria das aplicações práticas. Sendo recomendada mais para ter-se uma ideia de alto nível do que se trata determinado corpus.

2.3 Extração dos Aspectos Implícitos e questões a serem tratadas

Conjugado com as abordagens acima, é importante tratar também a questão da resolução de co-referência. Refere-se ao problema de determinar múltiplas expressões em um documento que referenciam a mesma entidade ou aspecto. Tem grande importância para análise de sentimentos Baseada em Aspectos, pois sentenças que antes eram desconsideradas por falta de reconhecimento do alvo da opinião, passam a ser inclusas na

análise, aumentando-se o percentual de *recall* no resultado final.

No trabalho de Ding et al (2010) é proposta uma abordagem para resolução da co-referência, na qual se tenta determinar a ligação de menções de entidades e aspectos com seus respectivos pronomes referenciadores, utilizando para isso técnicas de aprendizado supervisionado. Os autores levaram em consideração a ideia da consistência da opinião para análise de sentenças regulares e comparativas e tentaram determinar quais palavras de opinião são comumente associadas a determinadas entidades/aspectos.

Além dos aspectos explícitos, que são expressos como substantivos e locuções substantivas, existem os aspectos implícitos, que geralmente são representados por adjetivos e advérbios, mas que podem ser verbos também. Para que a análise de sentimentos tenha uma maior acurácia, é importante que haja um mapeamento dos aspectos implícitos para sua respectiva forma explícita.

Em um trabalho desenvolvido por Su et al (2008) foi proposto um método de clusterização para realizar tal mapeamento onde foi explorada a relação entre aspectos explícitos e palavras de opinião, formando pares de co-ocorrência em uma sentença. O mapeamento é obtido através de uma clusterização iterativa, formando um grupo de aspectos explícitos e um grupo de palavras opinativas. A cada iteração é calculada a similaridade intra grupo e entre os grupos. A associação é modelada por um grafo bipartido. Assim um aspecto e uma palavra opinativa são ligados se eles co-ocorrem em uma sentença. Quanto maior for a frequência de co-ocorrência, maior será o valor atribuído a ligação. Por fim as ligações que obtiverem os maiores valores correspondem ao mapeamento final.

O mapeamento de características implícitas ainda é um assunto pouco pesquisado e que demanda melhoramentos. Outra abordagem possível, porém custosa, é a criação de uma base de mapeadores de características em função de um domínio específico de forma manual, onde para cada palavra que indica uma característica implícita é associada a característica correspondente.

Feita a extração dos aspectos é necessário que, aspectos sinônimos sejam agrupados em uma categoria. E cada categoria formada deve representar um único aspecto. Essa é uma etapa importante para a precisão da análise de sentimentos e que pode ser realizada com o auxílio de dicionários. Porém o problema pode não ser resolvido por com-

pleto, pois muitos sinônimos são dependentes do domínio. Assim, duas palavras podem ser sinônimas em um domínio e não representarem sinônimos em outro domínio. Além disso, expressões podem não serem identificadas corretamente por dicionários e algumas que descrevem o mesmo aspecto podem não representar sinônimos.

Para lidar com tais problemas, diversas abordagens foram utilizadas, entre as quais estão: métricas de similaridade e métodos de aprendizado semi-supervisionados.

A seguir serão apresentados trabalhos encontrados na literatura que estão diretamente relacionados a este projeto.

2.4 Trabalhos Relacionados

Neste capítulo será apresentada uma visão resumida do processo de Extração de Características, serão apresentados os trabalhos diretamente relacionados a este projeto e, finalmente, serão apresentadas algumas propostas de modificação nas abordagens utilizadas no processo de extração de características que foram adotadas neste projeto e que diferem dos trabalhos relacionados.

2.4.1 Visão Sumarizada do processo de Extração de Características

Antes de serem abordados os trabalhos relacionados a este projeto, é apresentado um resumo do processo de Extração de Características. Na Figura 2.1 são apresentadas as etapas que compõem o processo:

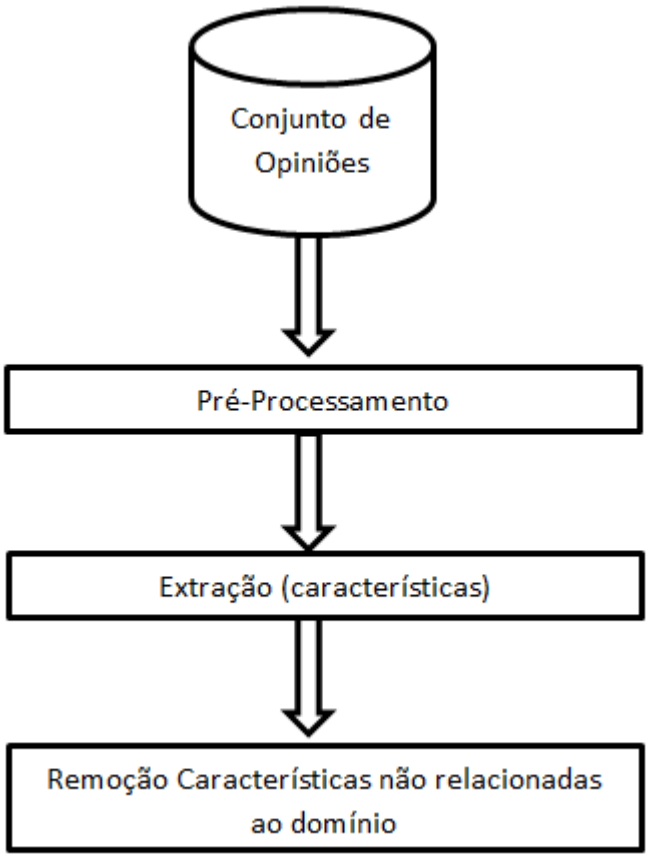

Figura 2.1: Etapas básicas do processo de extração de características

Primeiro tem-se um conjunto de opiniões, representando textos que contém mensagens opinativas cujas características presentes devem ser extraídas. Esse conjunto de opiniões geralmente é obtido através de um *web crawler*, que é um programa de computador que varre um conjunto de páginas na rede e que permite que se faça a coleta de alguma informação estipulada.

Tendo coletado todo texto opinativo é necessário que o mesmo passe por um pré-processamento. Esse processo é responsável por adequar os dados para a etapa de extração de características. Logo na etapa de pré-processamento algumas tarefas podem ser cumpridas, variando de acordo com cada projeto. As principais tarefas são:

- **identificação da subjetividade no texto de entrada:** essa tarefa permite que seja feita uma filtragem do texto a ser analisado, onde frases que contém somente fatos e não são opinativas são eliminadas. Com essa tarefa, apenas frases opinativas seriam analisadas no processo de extração de características, representando

portanto, uma tarefa de limpeza dos dados.

- **substituição de abreviações, correção gramatical e substituição de gírias:** na etapa de pré-processamento deve ser feito a rotulação de cada elemento presente no texto para sua devida classe gramatical através de um *PosTagger*. Elementos como abreviações, gírias e erros gramaticais atrapalham o desempenho do *PosTagger* na rotulação. Logo a tarefa de correção dos erros gramaticais, substituição de gírias por termos formais e substituição de abreviações pela forma extensa é importante, pois termos que forem mal classificados pelo *PosTagger* dada estas interferências, podem resultar em erros na etapa de extração das características.

- **rotulação:** esta tarefa é de extrema importância e seu êxito é fundamental para a etapa de extração das características. Logo para que o *POS Tagger* possa fazer um bom trabalho na rotulação dos elementos que compõem o texto, é necessário que o mesmo passe por um treinamento através de uma base rotulada, de preferência com o tipo de texto e domínio que será analisado, ou seja, se for uma análise de opiniões sobre o domínio "celular", o *POS Tagger* deveria preferencialmente ser treinado com textos opinativos, rotulados, cujo conteúdo seja referente a opiniões sobre o domínio celular .

Após a preparação dos dados, tem-se a etapa de extração das características propriamente dita. Nela são extraídos todos os elementos que representam o alvo de uma opinião. Alguns trabalhos, como por exemplo o de Lima (2011), propõem que nessa etapa, além das características, sejam extraídas também suas respectivas expressões opinativas. Representando assim um par, que corresponde a uma opinião. Isso ajudaria na etapa de Classificação de Sentimentos, uma vez que com a informação da extração estando neste formato, o contexto da opinião estaria preservado e expressões opinativas dependentes de contexto poderiam ser avaliadas com uma maior precisão.

Por fim, após as características ou os pares de opiniões serem extraídos, há uma etapa responsável pela eliminação dos pares de opinião não relacionados ao domínio. Após essa etapa, tem-se o processo completo de extração de características.

As etapas brevemente descritas anteriormente, representam as etapas básicas do

processo de Extração de Características. A seguir serão apresentados alguns trabalhos relacionados, cujas abordagens utilizadas pelos pesquisadores cumprem esta tarefa de extração.

2.4.2 Comparação entre os trabalhos relacionados

Foram encontradas na literatura três pesquisas, diretamente relacionadas ao trabalho desenvolvido neste projeto. Todas elas realizam as etapas básicas do processo de extração, porém, as abordagens utilizadas em cada projeto possuem algumas diferenças, que serão descritas para cada etapa. Bem como serão destacados os pontos de similaridade nas abordagens.

As três pesquisas se referem aos trabalhos de Hu et al (2004), Siqueira et al (2010) e Lima (2011). Segue na Tabela 2.1 um pequeno quadro resumo das abordagens e recursos utilizados em cada etapa dos trabalhos:

	Hu(2004)	Siqueira (2010)	Lima(2011)
Fonte das Opiniões	www.amazon.com WWW |net.com	www.ebit.com.br	www.google.com/prdhp
Domínio das opiniões	Produtos Duas câmeras digitais Um DVD Player Um MP3 Player Um celular	Serviços Prestados por empresas brasileiras	Produtos Filmes Celulares
Corpus	100 opiniões para cada subdomínio analisado	2200 opiniões Sendo 200 utilizadas na etapa de experimentos	100 opiniões para cada subdomínio analisado
Idioma	Inglês	Português	Inglês
POS-Tagger	NLP Processor	Tree Tagger	Tree Tagger
Extração das Características Frequentes	Regras de Associação	Contagem de Substantivos	Contagem de Substantivos
Extração de Palavras Opinativas	Proximidade entre substantivos / locuções substantivas com adjetivos	Proximidade entre substantivos com adjetivos	Regras gramaticais
Extração das características Infrequentes	Proximidade entre substantivos / locuções substantivas com adjetivos	Proximidade entre substantivos com adjetivos	Regras gramaticais
Mapeamento de Características Implícitas	Não realiza	A partir de uma base criada manualmente	A partir de uma base criada manualmente
Remoção das Características Não relacionadas	Feita na etapa de Características Frequentes (Poda)	PMI-IR	PMI-IR NGD

Tabela 2.1: Resumo das etapas e recursos utilizados por cada trabalho relacionado

Uma breve comparação dos projetos será realizada a seguir, por ordem de sequência das etapas. E por fim, será feita uma análise dos resultados obtidos, bem como será visto qual abordagem o projeto desenvolvido neste trabalho pretende utilizar visando

melhorar o processo de extração de características.

Base de Opiniões

O início do processo de Extração de Características requer um conjunto de opiniões como ponto de partida, na qual as características serão identificadas e extraídas.

No trabalho de Hu et al (2004), as opiniões foram coletadas dos seguintes sites: www.amazon.com e www.c|net.com, sendo obtidos comentários referentes a cinco produtos eletrônicos, sendo: duas câmeras digitais, um DVD player, um mp3 player e um celular (100 comentários para cada produto).

No trabalho de Siqueira et al (2010), a criação do corpus de opiniões foi realizada a partir de mensagens extraídas do site www.ebit.com.br, que contém a avaliação dos serviços prestados por diversas empresas. Extraiu-se 2200 opiniões para formar dois corpus. Um contendo 2000 opiniões utilizadas como base para adquirir conhecimento sobre o domínio e outra de 200 opiniões usada como uma base de validação para os experimentos.

No trabalho de Lima (2011), o corpus de opiniões foi formado por 100 opiniões do domínio de filmes e 100 opiniões do domínio de celulares. Todas retiradas do site www.google.com/prdhp, endereço que contém a opinião de vários consumidores com relação a uma grande variedade de produtos.

Portanto nos trabalhos de Hu et al (2004) e Lima (2011) o foco foi a Extração de Características de mensagens relativas à avaliação de produtos, enquanto no trabalho de Siqueira et al (2010) a Extração de Características se deu sobre um corpus cujo domínio é de avaliação de serviços prestados por empresas.

Pré-Processamento

Para que as Características pudessem ser extraídas dos textos opinativos, nos três trabalhos foi realizado um pré-processamento: segmentação das sentenças, remoção de *stopwords* e rotulação de cada termo presente no corpus com a sua correspondente classificação gramatical, com o auxílio de um *PosTagger*.

O *PosTagger* utilizado em cada projeto está identificado na Tabela 2.1.

Além das tarefas básicas do pré-processamento, no trabalho de Hu et al (2004),

os erros gramaticais foram tratados com a técnica de *Fuzzy Matching* e no trabalho de Lima (2011), são criadas bases de dados para substituir gírias, abreviações e contrações.

Extração das Características Frequentes

Para computar os Candidatos a Características Frequentes no trabalho de Hu et al (2004) foi utilizado a técnica de Regras de Associação. Onde se calculou os *itemsets* mais frequentes compostos por um grupo de palavras menor ou igual a três, pois características de produtos geralmente não contêm mais que três palavras. Para ser um item set frequente adotou-se um suporte mínimo de 1%.

Nos trabalhos de Siqueira et al (2010) e Lima (2011) a identificação dos Candidatos a Características Frequentes se deu a partir da contabilização da frequência dos substantivos onde o corte foi feito a partir de um *threshold* definido experimentalmente.

Poda dos Candidatos a Características Frequentes

Foi realizada somente no trabalho de Hu et al (2004). De acordo com sua pesquisa nem todos os Candidatos a Características Frequentes são genuínos. Existem aqueles não interessantes e redundantes. Assim o trabalho propõe dois tipos de poda para remover essas características indesejáveis.

Poda por Compacticidade

Em regras de associação o algoritmo não considera a posição de um item em uma transação. Porém em linguagem natural, palavras que aparecem juntas e em uma específica ordem em uma sentença são mais prováveis serem características significativas. Logo a poda por compacticidade visa eliminar aqueles candidatos a Características cujas palavras não aparecem próximas.

Assim, a distância entre duas palavras de um itemset extraído não pode ser maior que três palavras na sentença. Uma característica é dita compacta, se a regra anterior ocorre pelo menos em duas sentenças na base de opiniões. Um exemplo abaixo:

Eu procurei por uma boa câmera digital como essa por três meses.
Esta é a melhor câmera digital do mundo.
A câmera não possui um zoom digital.

Neste exemplo nas duas primeiras frases, as palavras câmera e digital respeitam a regra, tendo uma proximidade menor do que três palavras entre elas. Já na terceira frase essa regra da distância não é respeitada. Mas como no corpus aparecem pelo menos duas sentenças em que a distância entre as palavras estão dentro do limite da regra da distância, teremos neste exemplo "câmera digital" sendo considerada uma característica compacta, portanto, não passível de poda.

Poda de Redundância

Outro tipo de poda realizado no trabalho de Hu et al (2004) ocorre em substantivos que representam um subconjunto de locuções substantivas e que ocorram em uma frequência menor do que determinado suporte. Assim pode ser que, por exemplo, "vida de bateria" seja uma característica verdadeira para o domínio. Mas a palavra "vida", aparecendo isoladamente em outra sentença, pode não ser uma característica verdadeira, sendo podada caso a sua ocorrência esteja abaixo de um suporte mínimo definido.

Extração das Palavras Opinativas

Para a Extração dos Candidatos a Características Infrequentes é necessário que as palavras opinativas associadas aos Candidatos a Características Frequentes sejam extraídas.

Nos trabalhos de Hu et al (2004) e Siqueira et al (2010) essa extração é realizada considerando a presença de um adjetivo imediatamente anterior ou posterior ao Candidato a Característica Frequente. Essa abordagem leva em conta a proximidade entre características e palavras opinativas.

Já no trabalho de Lima (2011), a extração de palavras opinativas ocorre através de padrões que seguem regras gramaticais.

Extração das Características Infrequentes

Nos trabalhos de Hu et al (2004) e Siqueira et al (2010) a partir das palavras opinativas extraídas na etapa anterior, buscou-se por substantivos e locuções substantivas que ocorriam no corpus nas proximidades das palavras opinativas. Caso fosse encontrado um substantivo ou locução substantiva e esse não pertencesse à lista de Candidatos a Características Frequentes, então ele era adicionado à lista de candidatos à características infrequentes, que são aquelas palavras que mesmo ocorrendo com menor frequência, podem representar características importantes para o domínio analisado.

Já no trabalho de Lima (2011), a extração dos Candidatos a Características Infrequentes foi realizada também considerando as palavras opinativas associadas aos Candidatos a Características Frequentes, porém foi realizada utilizando padrões que seguem regras gramaticais ao invés da distância entre substantivos e locuções substantivas em relação às palavras opinativas.

Mapeamento de Características Implícitas

Nos trabalhos de Siqueira et al (2010) e Lima (2011) os mapeamentos foram realizados a partir de uma base de mapeamentos criada de forma manual. No trabalho de Hu et al (2004) essas características foram desconsideradas, uma vez que ocorrem em menor quantidade em relação às características explicitas.

Remoção de Características não relacionadas ao domínio

No trabalho de Hu et al (2004), essa etapa foi realizada somente para os Candidatos a Características Frequentes e já foi descrita anteriormente, representando a etapa de poda.

Já nos trabalhos de Siqueira et al (2010) e Lima (2011), essa remoção é mais abrangente, considerando todos os Candidatos a Características: Frequentes, Infrequentes e Implícitas.

No trabalho de Siqueira et al (2010), foi utilizada uma medida probabilística denominada *Pointwise Mutual Information - Information Retrieval (PMI-IR)* para calcular o grau de similaridade entre o Candidato a Característica e o domínio em questão e assim remover as características não relacionadas ao domínio baseado em um *threshold*

experimental.

No trabalho de Lima (2011), além da medida *(PMI-IR)*, foi utilizada outra medida probabilística para o cálculo de similaridade denominada *Normalized Google Distance (NGD)*. Nos resultados comparativos foi considerada mais efetiva, melhorando a medida de precisão [5] nos resultados finais.

Vale ressaltar que, apesar de todos os trabalhos extraírem palavras opinativas, o trabalho de Lima (2011) mantém o par (Característica, Expressão Opinativa) como resultado de saída final, ou seja, a saída é dada em pares que representam o contexto da opinião.

A etapa de remoção dos Candidatos a Características não relacionados ao domínio representa a última etapa do processo de Extração de Características, o qual resulta em uma lista contendo somente características verdadeiras (ou pares de características e expressões opinativas associadas) e importantes para o domínio analisado.

Resultados

Na Tabela 2.2 os resultados obtidos no trabalho de Hu et al (2004):

Produtos	Número de Caracte-rísticas	Características Frequentes (regras de associação)		Poda por Compacticidade		Poda por Redundância		Identificação das Características Infrequentes	
		Cobertura	Precisão	Cobertura	Precisão	Cobertura	Precisão	Cobertura	Precisão
Câmera digital 1	79	0.671	0.552	0.658	0.634	0.658	0.825	0.822	0.747
Câmera digital 2	96	0.594	0.594	0.594	0.679	0.594	0.781	0.792	0.710
Celular	67	0.731	0.563	0.716	0.676	0.716	0.828	0.761	0.718
MP3 Player	57	0.652	0.573	0.652	0.683	0.652	0.754	0.818	0.692
DVD Player	49	0.754	0.531	0.754	0.634	0.754	0.765	0.797	0.743
Média	69	0.68	0.56	0.67	0.66	0.67	0.79	0.80	0.72

Tabela 2.2: Resultados obtidos no trabalho de Hu et al (2004)

Foi avaliado o resultado de cada etapa do algoritmo a partir de mensagens opinativas relativas aos produtos eletrônicos já citados anteriormente.

[5] Métrica utilizada para avaliação do processo de extração de características. Refere-se ao total de características verdadeiras extraídas em relação a todas características extraídas.

Concluiu-se que só com a etapa de extração de Características Frequentes se obtém resultados pouco satisfatórios. Uma explicação é a grande quantidade de características extraídas que não são significativas para o domínio e a não extração de outras características que são importantes para o domínio, mas não são frequentes, logo não são extraídas nessa etapa.

Com as etapas subsequentes de poda, observou-se que houve uma melhora significativa na medida de precisão, ou seja, muitas características não significativas para o domínio foram descartadas. Nessas etapas a medida de cobertura[6] permaneceu praticamente constante.

Por fim, a etapa de Extração de Características Infrequentes resultou em uma melhora significativa na medida de cobertura e um pequeno decréscimo na medida de precisão. Demonstrando que a Extração de Características Infrequentes aumentou o número de Características extraídas verdadeiras, mas extraiu também algumas características falsas.

Para esse experimento os pesquisadores obtiveram em média 80% de cobertura e 72% de precisão na tarefa de Extração de Características, o que foi considerado promissor, validando as técnicas a serem utilizadas para questões práticas do mundo real.

A seguir,na tabela 2.3 são apresentados os resultados obtidos no trabalho de Siqueira et al (2010):

Etapas	Precisão	Cobertura
1	28.09%	24.91%
1+2	29.12%	53.96%
1+2+3	50.62%	92.08%
1+2+3+4	77.24%	90.94%

Tabela 2.3: Resultados obtidos no trabalho de Siqueira et al (2010)

Foi avaliado o resultado de cada etapa do algoritmo. Para a primeira etapa que representa a extração das 3% características mais frequentes obteve-se baixos resultados de precisão e cobertura.

Na segunda etapa que representa a extração das características infrequentes so-

[6]Métrica utilizada para avaliação do processo de extração de características. Refere-se ao total de características verdadeiras extraídas em relação a todas características verdadeiras do corpus.

madas as características frequentes já obtidas anteriormente, resultaram em: aumento significativo de cobertura e um pequeno aumento na precisão.

Na terceira etapa que representa os resultados da segunda adicionados ao mapeamento das características implícitas, obteve-se uma melhora significativa tanto na precisão como na cobertura.

Na quarta e última etapa, que representa a etapa de filtragem das características obteve-se um aumento substancial na medida de precisão, o que representa que características irrelevantes foram descartadas assim como um pouco de relevantes, que fez com que houvesse uma pequena queda na medida de cobertura.

Segue nas Tabelas 2.4 e 2.5 os parâmetros de teste e resultados obtidos no trabalho de Lima (2011):

Configuração	Domínio do Corpus	*Threshold* Frequência	*Threshold* Relevância
1.1	Aparelhos Celulares	5%	0.3
1.2	Aparelhos Celulares	5%	0.5
1.3	Aparelhos Celulares	10%	0.3
1.4	Aparelhos Celulares	10%	0.5
2.1	Filmes	5%	0.3
2.2	Filmes	5%	0.5
2.3	Filmes	10%	0.3
2.4	Filmes	10%	0.5

Tabela 2.4: Parâmetros de teste do trabalho de Lima (2011)

	Etapa 1			Etapa 2			Etapa 3		
Configuração	P	C	FM	P	C	FM	P	C	FM
1.1	0.809	0.239	0.369	0.625	0.704	0.662	0.666	0.704	0.684
1.2	0.809	0.239	0.369	0.625	0.704	0.662	0.725	0.633	0.676
1.3	0.685	0.338	0.452	0.602	0.746	0.666	0.641	0.732	0.684
1.4	0.685	0.338	0.452	0.602	0.746	0.666	0.696	0.647	0.671
2.1	0.764	0.203	0.320	0.518	0.640	0.573	0.493	0.625	0.551
2.2	0.764	0.203	0.320	0.518	0.640	0.573	0.631	0.562	0.595
2.3	0.689	0.312	0.430	0.558	0.750	0.639	0.610	0.734	0.666
2.4	0.689	0.312	0.430	0.558	0.750	0.639	0.682	0.671	0.677
Média	**0,736**	**0,273**	**0,392**	**0,575**	**0,71**	**0,635**	**0,643**	**0,663**	**0,649**

Tabela 2.5: Resultados do trabalho de Lima (2011)

Foi avaliado o resultado de cada etapa do algoritmo. Neste trabalho o mapeamento das Características Implícitas ocorre junto com a Extração dos Candidatos a Características Frequentes e Infrequentes. Para a etapa 1 que representa a Extração dos Candidatos a Características Frequentes obtiveram-se os índices mais altos de precisão e mais baixos de cobertura. A explicação para tal foi que com a Extração dos Candidatos a Características Frequentes o esperado é que se obtenha uma proporção pequena de características falsas, pois características frequentes tendem a serem genuínas, mas elas representam uma pequena fração de todas as características verdadeiras.

Na segunda etapa, que representa o resultado da primeira somado a extração dos candidatos a características infrequentes, percebeu-se uma redução na precisão, pois novas características foram acrescidas, sendo estas verdadeiras e falsas, e consequentemente houve um aumento na medida de cobertura por conta desse adicional de novas características.

Na etapa 3, que representa a filtragem das características com base na relevância com o domínio, observou-se uma pequena redução da cobertura e um aumento na precisão. Isso ocorreu, pois é descartada grande parte dos pares que contém características falsas e eventualmente alguns pares que contém características verdadeiras.

Observou-se também que os resultados do domínio de aparelhos celulares foram melhores, comparados aos resultados do domínio de filmes, devido a menor convergência do conjunto de características deste domínio, variando de acordo com cada filme.

2.4.3 Considerações Finais

O melhor resultado obtido dos três trabalhos foi o de Siqueira et al (2010) apresentando valores de 77,24% de precisão e 90,94% de cobertura para o corpus analisado. Porém ao acessar o site em que as mensagens foram retiradas, pode-se perceber que as mesmas eram bem simples, na maioria das vezes não passando de uma linha de comprimento. Este fato pode ter facilitado o processo de Extração de Características, possibilitando se alcançar valores tão altos de acurácia.

O segundo melhor resultado foi encontrado no trabalho de Hu et al (2004), apresentando como resultado final 72% de precisão e 80% de cobertura em média. Ao acessar

as mensagens utilizadas na extração de características, verificou-se que são mensagens bem variadas: muitas mensagens curtas e médias e um pouco de mensagens longas. Logo parece ser um bom resultado considerando a complexidade das sentenças.

Por fim, o trabalho de Lima (2011) foi o que apresentou os piores resultados, 64% para precisão e 66% para cobertura em média. Ao analisar a estrutura das mensagens utilizadas verificou-se que as mensagens são bem variadas, mas são compostas em sua maioria por mensagens médias e longas. Talvez isto tenha contribuído para resultados de acurácia mais baixos, uma vez que sentenças mais longas tendem a ser mais complexas e, portanto, mais difíceis de lidar.

Pode-se observar que os trabalhos são complementares. Por exemplo, apesar do trabalho de Hu et al (2004) ter apresentado bons resultados finais, nele não são filtrados todos os Candidatos a Características, somente os frequentes. E também não é abordado a Extração das Características Implícitas como nos trabalhos de Siqueira et al (2010) e Lima (2011). Muito provavelmente uma adaptação em sua metodologia melhoraria ainda mais os resultados finais.

Acaba sendo difícil a comparação entre os trabalhos pelo fato de cada um deles ter utilizado bases de opiniões diferente e módulos diferentes. Por exemplo, no trabalho de Siqueira et al (2010) a extração de palavras opinativas é realizada baseada na proximidade dos Candidatos a Características Frequentes com o adjetivo modificador. Assim palavras opinativas se limitam a serem adjetivos. Enquanto no trabalho de Lima (2011), a extração de palavras opinativas é baseada em padrões de acordo com regras gramaticais. Pode-se desta maneira extrair palavras opinativas de uma maneira mais abrangente, podendo estar representadas na forma de substantivo, verbo, além da forma mais comum que é adjetivo. Mas fica difícil saber qual das duas abordagens foi mais eficiente, pois as etapas de cada projeto, não são totalmente correspondentes. Teoricamente a abordagem de Lima (2011) é mais abrangente para extrair palavras opinativas e consequentemente Extrair Características Infrequentes, que dependem das palavras opinativas. Porém como visto o seu resultado foi pior que o de Siqueira et al (2010), que por sua vez trabalhou com mensagens mais simples, o que explicaria a boa acurácia.

Enfim, os três trabalhos apresentam boas abordagens. E apesar de serem difíceis

de serem comparados, representam um bom ponto de partida para o desenvolvimento de um protótipo de Extração de Características e Palavras Opinativas, visto que suas abordagens demonstraram-se efetivas para resolução do problema.

Tendo como base as abordagens anteriores, neste projeto decidiu-se fazer uma adaptação das técnicas vistas. Para a etapa de extração das características, foi visto que alguns projetos utilizaram a abordagem de extração por proximidade entre substantivos e adjetivos, o que parece ser uma boa técnica, visto que a maioria das opiniões é composta por um adjetivo modificando um substantivo. Porém, essa abordagem tem a limitação de ter como expressão opinativa apenas adjetivos.

A outra abordagem de extração utilizada foi por regras gramaticais, que tem a limitação de serem rígidas, ou seja, se não houver a correspondência total da classe gramatical do texto com a regra, a característica não é extraída. Porém ela é mais abrangente com relação à extração de tipos de opiniões, ou seja, com essa abordagem uma expressão opinativa pode além de ser um adjetivo, assumir outras classes gramaticais, como advérbio, verbo ou substantivo. Assim como a própria característica extraída, pode assumir um formato mais complexo, como uma locução, que provavelmente não seria extraída pela regra da proximidade.

Neste projeto, a abordagem escolhida foi unir as duas técnicas anteriores, buscando aproveitar os pontos fortes de cada abordagem e com isso tornar a etapa de extração mais completa. Foi adotada também a extração das características, junto com suas respectivas palavras opinativas, de forma a preservar o contexto da opinião e aumentar a precisão da etapa de classificação de sentimentos, principalmente com relação à classificação de expressões opinativas dependentes de contexto.

Ainda foi adotada uma pequena modificação, onde as características são extraídas em sua totalidade e somente no final do processo que se escolhe o percentual mais frequente que se deseja considerar. Com isso as sub etapas de extração de características frequentes e infrequentes são fundidas em apenas uma etapa. Assim, evita-se a limitação de um candidato à característica infrequente ser formado apenas a partir das expressões opinativas dos candidatos a características frequentes. Objetivando com isso aumentar o número de pares extraídos.

Essa foi apenas uma sumarização da abordagem adotada neste projeto, identificando pontos que se diferem das adotadas em outros projetos. A seguir será detalhada a metodologia utilizada nesse trabalho.

3 Abordagem adotada para extração de características em textos de língua portuguesa

A abordagem utilizada nesse projeto para o desenvolvimento do protótipo, é baseada em métodos estatísticos e processamento de linguagem natural. Na Figura 3.1 é apresentada uma visão geral da arquitetura do sistema:

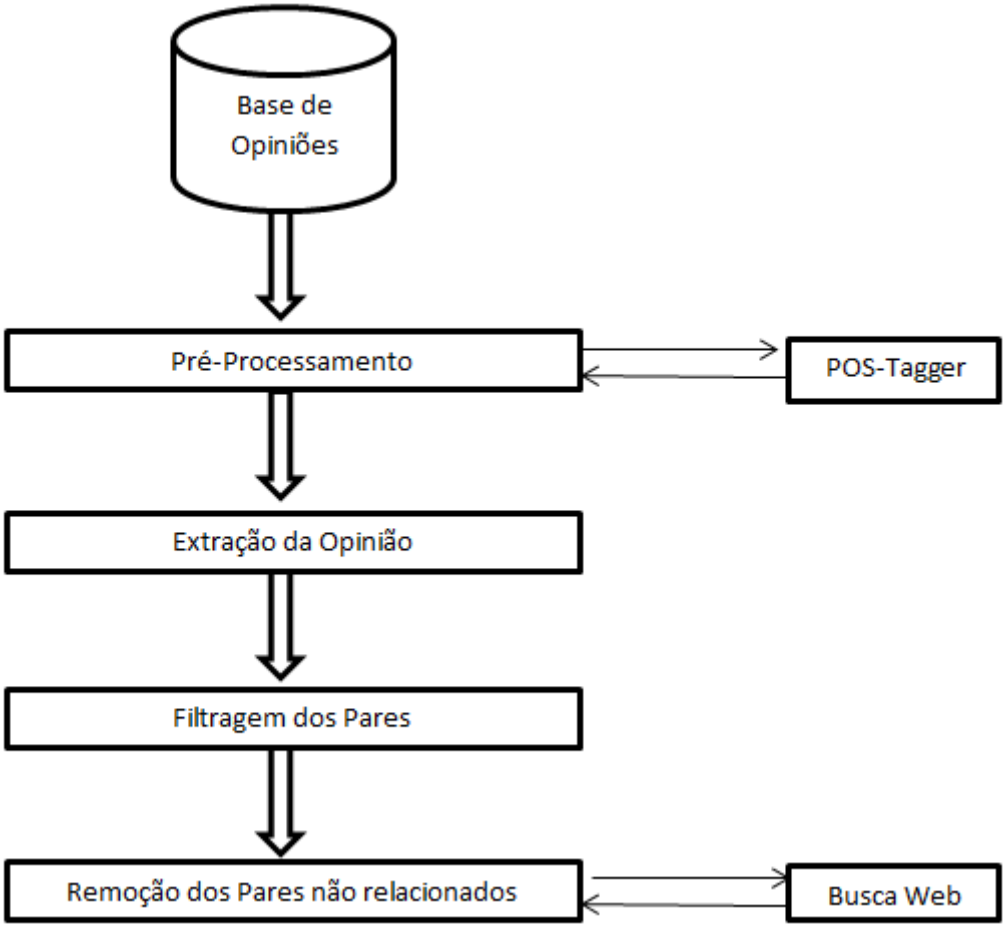

Figura 3.1: Arquitetura do Protótipo

Essa arquitetura é bem semelhante àquela mostrada no Capítulo 2, que demonstra uma arquitetura básica de um processo de extração de características. Mas como etapa adicional tem-se a filtragem dos pares, que foi adicionada devido à abordagem adotada de realizar a extração das características utilizando a técnica que considera a proximidade

entre substantivos e adjetivos e também a técnica que envolve regras gramaticais. Essa utilização conjunta das técnicas acaba por gerar duplicatas, logo foi necessário adicionar um módulo para filtrar esses casos, que será descrito nas próximas seções, assim como os todos os detalhes da metodologia adotada neste trabalho.

3.1 Pré-Processamento

O primeiro módulo, pré-processamento, tem a função de preparar o corpus de entrada para o processo de extração das opiniões.

Para esse projeto, nesse módulo são realizadas somente duas operações. A primeira refere-se à tokenização. Esta operação é realizada com auxílio da biblioteca Apache Open NLP, onde cada unidade que constitui o texto (cada palavra ou pontuação) é identificada e separada em tokens. Conforme o exemplo abaixo:

A qualidade de imagem é ótima. (Corpus de entrada)
"A", "qualidade", "de", "imagem", "é", "ótima", "."(Corpus após a tokenização)

Neste exemplo a partir do corpus de entrada, seriam retornados sete tokens.

A partir do corpus tokenizado é realizada a segunda e última operação, que representa a etiquetagem dos tokens. Mais uma vez, é utilizada a biblioteca Apache Open NLP, onde com o auxílio de um *POS Tagger*, ferramenta de processamento de linguagem natural, é realizada a leitura de todos os tokens que compõem as mensagens com a rotulação de suas devidas classes gramaticais. A seguir um exemplo:

"A", "qualidade", "de", "imagem", "é", "ótima", "."

Acima se tem os tokens identificados do corpus de entrada. E abaixo as respectivas etiquetas, ou seja, as classes gramaticais dos tokens analisados.

“artigo” , “substantivo” , “preposição” , “substantivo” , “verbo” , “adjetivo” , “pontuação”

Após esta etapa ser cumprida, a etapa de extração de opiniões é realizada conforme a próxima seção.

3.2 Extração de Opiniões

O módulo de Extração de Opiniões utilizará como entrada o conjunto de textos contendo opiniões devidamente tokenizado e com suas respectivas etiquetas gramaticais. O processo de Extração de Opiniões é composto por duas etapas: extração dos pares candidatos à opinião por proximidade e extração dos pares candidatos à opinião por regras gramaticais.

3.2.1 Extração dos pares por proximidade

Nesta etapa são extraídos os pares candidatos a opiniões, onde as características são representadas por substantivos ou locuções substantivas, e as expressões opinativas são representadas apenas por adjetivos.

Aqui se segue a heurística de que características são vinculadas a expressões opinativas com a menor distância. Exemplo:

O tempo de bateria é ruim, mas o celular é muito bom.

Nesse exemplo seriam extraídos os pares:

(tempo de bateria -> ruim) (celular -> bom)

Note que neste nesse exemplo o segundo par extraído está incompleto, o certo seria (celular -> muito bom). Isto ocorre dada a limitação do método de ser considerada expressão opinativa apenas adjetivos. Para corrigir essa falha e evitar que a opinião seja extraída de forma incompleta a próxima etapa utiliza outra estratégia de extração, por regras gramaticais, que irá complementar a extração desta etapa. E na etapa de filtragem que será vista mais a frente, o par que traz a opinião incompleta é eliminado.

Apesar de ser um método limitado, ele é importante, pois independe de padrões fixos, logo se uma opinião não é extraída pelo método de regras gramaticais, seja por inexistência de alguma regra ou por erros na rotulagem dos tokens, pode ser que seja extraída pelo método de proximidade.

3.2.2 Extração dos pares por regras gramaticais

Após a identificação e extração dos candidatos a opiniões pela primeira heurística que leva em consideração a proximidade entre características e palavras opinativas, percebe-se que a abordagem não detecta todas as opiniões do corpus. Isto porque ela é muito simples, apesar de considerar o caso mais comum (adjetivos sendo expressões opinativas). Necessita-se assim de outra abordagem que identifique os outros tipos de opiniões.

Portanto, para complementação da extração anterior, nesta etapa a extração da opinião se dá por meio de regras gramaticais. Onde, além de adjetivos, expressões opinativas também podem ser extraídas estando em outras classes gramaticais como: advérbios, substantivos, verbos e expressões compostas por vários termos. Assim como se tem a possibilidade de extração de opiniões onde características não são explicitadas na forma de substantivo ou locuções substantivas. Essas opiniões estão implícitas. Nesse caso de haver características implícitas, a característica é nomeada "Geral" e representa a própria entidade. Exemplo:

É fantástico.
Opinião extraída: ("Geral", Fantástico")

Como o sujeito não é determinado, o alvo da opinião, ou seja, a característica em questão é desconhecida. Assim assume-se que nesses casos a característica representa a própria entidade analisada. Essa relaxação é feita, pois é verdadeira na maioria dos casos na análise de textos opinativos. É na verdade uma forma bem simplificada de realizar a resolução de correferência.

A construção das regras de opiniões foi realizada levando-se em consideração as regras da gramática portuguesa e inferidas a partir da observação do corpus analisado. Para cada regra, quatro variações foram criadas:

- **com sujeito determinado e negação**

Exemplo: O celular não é fácil de usar.
Regra gramatical: substantivo + advérbio + verbo + adjetivo + preposição + verbo
Par extraído: (celular -> não é fácil de usar)

- **com sujeito determinado e afirmação**

Exemplo: O celular é fácil de usar.
Regra gramatical: substantivo + verbo + adjetivo + preposição + verbo
Par extraído: (celular -> fácil de usar)

- **com sujeito indeterminado e negação**

Exemplo: Não é fácil de usar.
Regra gramatical: advérbio + verbo + adjetivo + preposição + verbo
Par extraído: (Geral -> não é fácil de usar)

- **com sujeito indeterminado e afirmação**

Exemplo: É fácil de usar.
Regra gramatical: verbo + adjetivo + preposição + verbo
Par extraído: (Geral -> fácil de usar)

Com essa etapa aumenta-se não só o número de candidatos a pares de opinião extraídos, mas também se consegue extrair opiniões mais significativas semanticamente. Todas as regras de opinião elaboradas neste trabalho estão descritas no Apêndice.

3.3 Filtragem dos Pares

Após a extração dos candidatos à opinião pelas duas técnicas anteriores, podem ocorrer colisões, ou seja, opiniões que são extraídas representam duplicatas. Para resolver esse problema, criou-se um filtro. Vale destacar que candidatos à opinião em duplicata não necessariamente representam candidatos idênticos. Por exemplo:

Opinião (iPhone -> bom)
Opinião (iPhone -> muito bom)

Neste exemplo temos o primeiro candidato à opinião obtido pela etapa de extração dos pares por proximidade, já o segundo candidato à opinião, é obtido através da etapa de extração dos pares por regras gramaticais. Apesar de não serem iguais, representam a opinião do mesmo trecho no *corpus*. Logo, se considera que os candidatos a opiniões estão duplicados. Assim o filtro eliminaria o primeiro candidato a opinião, mantendo o

segundo que no caso representaria o candidato cuja opinião possui mais representatividade semântica.

Um objeto Opinião neste projeto é representado por uma quádrupla, onde o primeiro campo representa uma característica, o segundo a posição da característica, o terceiro a expressão opinativa e o quarto a posição da expressão opinativa. Quando uma Opinião não tem um sujeito determinado, considera-se que a opinião é sobre a entidade e denomina-se tal característica como "Geral" e sua posição recebe a posição zero simbolicamente.

O filtro leva em consideração todas as variações dos elementos que compõem uma opinião. Os pares exemplificados abaixo, não representam todos os pares que são gerados pelo algoritmo e sim os pares que são afetados diretamente pela ação da parte do filtro discutida. Os números entre parênteses representam as posições dos termos. Assim temos os seguintes casos:

- **Posição entre as características são diferentes e entre as expressões opinativas são iguais. E nenhuma das características é "Geral".** Exemplo:

A tela é boa mas a bateria poderia ser muito melhor.

  Neste exemplo pela estratégia da proximidade, seriam extraídos os pares:

Tela (1) -> boa (3) (distância = 2)
Bateria (6) -> boa (3) (distância = 3)

  Nesta situação, o filtro somente mantém o par que apresentar a menor distância (posição da característica – posição da expressão opinativa) em módulo. Logo, o par (bateria -> boa), seria eliminado permanecendo somente o par (tela -> boa), que representa a menor distância. Neste exemplo o par correto para extração da característica "Bateria" seria (bateria -> poderia ser muito melhor). Pela etapa de extração por proximidade se perderia essa opinião. Mas ela é extraída pela etapa da extração por regras gramaticais. Logo as etapas se complementam.

- **Posição entre as características são iguais e entre as expressões opinativas são iguais. E nenhuma das características é "Geral".** Exemplo:

Produto muito bom.

A partir das duas estratégias de extração seriam extraídos os pares:

Produto (0) -> bom (2)
Produto (0) -> muito bom (2)

Aqui o filtro adiciona a expressão de opinião com maior significado semântico e elimina as demais. Neste caso (Produto -> bom) seria eliminado, permanecendo (produto -> muito bom).

Por conveniência adotou-se usar as seguintes regras: Para a característica formada por uma locução substantiva, a posição do termo é sempre a última. Exemplo:

A qualidade de imagem é muito boa e a bateria apesar de tudo razoável.
Qualidade de imagem (1) -> boa (6) (considerando ordem normal -> distância =5)
Qualidade de imagem (3) -> boa (6) (considerando regra adotada -> distância = 3)
Bateria (9) -> razoável (13)

Se não fosse adotada tal regra nesse exemplo, pela heurística da proximidade, os pares identificados seriam:

Qualidade de imagem (1) -> boa (6) (Distância =5)
Qualidade de imagem (1) -> razoável (13) (Distância =12)
bateria (9) -> boa (6) (Distância =3)
bateria (9) -> razoável (13) (Distância =4)

Para característica qualidade de imagem o par com menor distância, portanto extraído, seria (qualidade de imagem (1) -> boa (6)) e o par com menor distância para a característica bateria seria (bateria (9) -> boa (6)). Como esses dois pares possuem a mesma posição de expressão opinativa e posições de características diferentes, pelo filtro o par com menor distância que permaneceria, que seria o (bateria (9) -> boa (6)), logo representando um par errado. Com a regra adotada esse tipo de caso fica mais difícil de acontecer. A distância nesse caso seria 3 a mesma encontrada em (bateria (9) -> boa (6)) e em caso de empate, pelo algoritmo, só o

primeiro caso permanece. Que representa no exemplo o par correto: (Qualidade de imagem (3) -> boa (6))

Na expressão opinativa a regra adotada é que a posição opinativa é sempre a posição do adjetivo quando houver. Exemplo:

A imagem é muito boa.
Imagem (1) -> boa (4)
Imagem (1) -> muito boa (4)

Isso facilita o filtro, pois é uma forma fácil de saber que ambas as expressões representam a mesma opinião.

- **Posição entre as características são iguais e entre as expressões opinativas são diferentes. E nenhuma das características é "Geral".** Exemplo:

Produto bom, bonito e barato.

Extrações:

Produto (0) -> bom (1) (Regra proximidade) I
Produto (0) -> bom (1) (Regras gramaticais) II
Produto (0) -> bonito (3) (Regras gramaticais) III
Produto (0) -> barato (4) (Regras gramaticais) IV

Essa parte do filtro trata os casos II, III, IV. Todos os candidatos à opinião em uma enumeração são adicionados.

- **Posição entre as características são diferentes e entre as expressões opinativas são diferentes. E nenhuma das características é "Geral".** Exemplo:

A qualidade de imagem é muito boa.

Extrações:

Qualidade de imagem (3) -> boa (6) I
Qualidade de imagem (1) -> muito boa (4) II

O filtro compara os campos característica. Se um campo for substring de outro, o par que contiver o maior comprimento para a característica é mantido e os outros que são substrings são eliminados. Caso os campos não sejam substrings um do outro, todos são adicionados.

Essa parte do filtro possibilita que mesmo havendo erro na adoção das regras de posição, ou seja, o par II deveria ter tido as seguintes posições (Qualidade de imagem (3) -> é muito boa (6)), o par correto seja escolhido.

- **Os campos de característica comparados possuem ambos o valor "Geral".** Exemplo:

Poderia ser melhor.

Extrações:

Geral (0) -> melhor (2)
Geral (0) -> poderia ser melhor (2)

O filtro compara os campos de expressão opinativa. Se um campo for substring de outro, o par que contiver o maior comprimento para a expressão opinativa é mantido e os outros que são substrings são eliminados. Caso os campos não sejam substrings um do outro, todos são adicionados.

- **Um campo característica possui o valor "Geral" e o outro é diferente de "Geral".** Exemplo:

Funciona de forma consistente.

Extrações:

Geral (0) -> funciona de forma consistente (3)
Forma (2) -> consistente (3)

O filtro compara os campos de expressão opinativa. Se um campo for substring de outro, o par que contiver o maior comprimento para a expressão opinativa é mantido e os outros que são substrings são eliminados. Caso os campos não sejam substrings um do outro, todos são adicionados.

Este processo de filtragem dos pares evita que a listagem final de candidatos à opinião extraídos contenha duplicatas e faz com que estas quando ocorrem ao longo do processo de filtragem, sejam substituídas por um único candidato a opinião, o qual é o mais representativo semanticamente entre as duplicatas. Por receber os candidatos à opinião ordenados, o processo garante que esses pares finais obtidos estejam na ordem real em que as opiniões são lidas no corpus, o que facilita a checagem dos resultados finais por um ser humano.

3.4 Remoção dos pares não relacionados

Nas etapas anteriores foram extraídos os candidatos a opiniões. Porém, uma parte desses candidatos não contém de fato características verdadeiras para um determinado domínio. É necessário que seja feita uma remoção de tais candidatos para que seja definida uma listagem final, contendo apenas as opiniões relacionadas ao domínio pesquisado, sendo as demais descartadas. Por exemplo:

O vendedor foi gentil e nos deixou ver todos os modelos de celular da loja.

Assuma que o domínio pesquisado seja sobre celular e que o seguinte candidato à opinião seja extraído pelo processo:

(vendedor, gentil)

A característica "vendedor" não representa uma característica importante para o domínio "celular", apesar de ser extraída. Assim necessita-se de alguma técnica para que esses candidatos à opinião falsos sejam eliminados.

Neste trabalho a técnica utilizada para realização da remoção dos pares obtidos não relacionados ao domínio pelas etapas anteriores é baseada na medida *NGD* (*Normalized Google Distance*). Essa técnica foi desenvolvida no trabalho de Cilibrasi et al (2007) e baseia-se na complexidade de *Kolmogorov* e na distância normalizada da informação.

Distância Google é uma medida de similaridade semântica derivada do número de retornos obtidos pelo motor de busca do Google a partir de um conjunto de palavras

chave. O cálculo da Distância Google Normalizada entre dois termos x e y é:

$$NGD(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logM - min\{logf(x), logf(y)\}} \quad (3.1)$$

Sendo:

M: é o número total de páginas indexadas pelo Google.

f(x) e f(y): são os números de páginas retornadas para cada termo pesquisado x, y respectivamente.

f(x,y): é o número de páginas retornadas em que os termos x e y aparecem no mesmo documento.

Se os termos pesquisados nunca ocorrem juntos em uma página, mas possuem ocorrências separadamente, a distância entre eles é infinita. Já se os termos pesquisados sempre ocorrem juntos, a distância entre eles é zero.

Assim, com o cálculo da distância *NGD* é possível calcular a relevância das características extraídas com relação ao domínio analisado e assim filtrar os pares de opiniões candidatos, resultando em uma lista das opiniões verdadeiras como resultado final do protótipo. Sendo o valor *NGD*, utilizado como referência para a filtragem definido experimentalmente e passado como um parâmetro para o sistema.

3.5 Conclusões

Neste capítulo foi apresentada a metodologia utilizada para extração da opinião, no formato de pares (características, palavras opinativas), que como já foi explanado, visa melhorar a precisão da etapa de classificação de sentimentos, uma vez que o contexto da opinião é preservado, logo expressões opinativas dependentes de contexto poderão receber uma valoração adequada. Pelo fato de a expressão opinativa ser composta não somente por adjetivos, e sim por elementos de várias classes gramaticais, como advérbios, substantivos, verbos e locuções, pode-se melhorar também a precisão do valor que a expressão de sentimento recebe no processo, que agora poderá ser de acordo com a intensidade dos elementos que compõem uma expressão opinativa, não se limitando a apenas um elemento de classe gramatical representada por um adjetivo.

Pelo fato da etapa de extração de opinião adotar duas técnicas, a que considera a proximidade entre substantivos e adjetivos e a realizada através de regras gramaticais, espera-se que os pares de opiniões extraídos sejam compostos por características e expressões opinativas mais ricas semanticamente, pois com regras gramaticais se consegue extrair esses elementos em formatos mais complexos e completos. E ainda, que a medida de cobertura, ou seja, o retorno dos pares de opiniões corretos em relação aos existentes no corpus analisado, aumente, pois com a utilização da técnica por proximidade, casos que não forem cobertos pelas regras gramaticais poderão ter a chance de serem extraídos, sendo assim uma etapa complementar.

Foram vistas aqui todas etapas e abordagens utilizadas para o desenvolvimento do protótipo proposto neste trabalho. No próximo capítulo serão apresentados os procedimentos de teste para validação do artefato que foi desenvolvido bem como os resultados obtidos.

4 Avaliação

Neste capítulo serão apresentados os procedimentos de testes realizados neste projeto. Antes de detalhar os testes, primeiro é descrito a criação dos *benchmarks*.

4.1 *Benchmarks* utilizados nos testes

Não foram encontrados para realização de testes do protótipo *benchmarks* de mensagens opinativas em língua portuguesa. Dada essa dificuldade, foram elaborados neste projeto duas bases de teste. Uma contendo textos opinativos referentes ao domínio de produtos e outra ao domínio de serviços.

O corpus relativo ao domínio de produtos foi extraído do site www.google.com[7], onde é encontrado grande número de avaliações referentes a diversos produtos. Para criação do corpus foram consideradas opiniões referentes ao aparelho celular "iPhone". A composição final da base foi formada por mensagens com tamanhos bem variados, sendo em sua maioria curtas e médias. As mensagens coletadas estavam originalmente no idioma inglês e foram traduzidas, e os erros gramaticais foram corrigidos, assim como a pontuação.

Já o corpus relativo a serviços, foi extraído do site www.ebit.com.br, que contém avaliações feitas por consumidores em relação a serviços prestados por diversas empresas. Foi composto em sua maioria por mensagens curtas e médias, sendo os erros gramaticais corrigidos, mas a pontuação se manteve original.

Para cada uma das bases de testes formadas, foi efetuada a extração de todos os pares de opiniões de forma manual, formando dois modelos finais que foram usados como gabarito para conferência das saídas do algoritmo nos testes realizados.

[7] www.google.com/prdhp

4.2 Procedimentos de Testes

Neste trabalho, a etapa de testes visa analisar dois aspectos que influenciam diretamente a acurácia do processo de extração de características (pares de opiniões), que são: o *POS Tagger* utilizado para a rotulação dos *tokens* e a abordagem adotada para a etapa de extração da opinião.

Nas subseções seguintes serão detalhados os procedimentos paras as análises.

4.2.1 Avaliação da influência do *POS Tagger* no processo de extração de Características(opiniões)

Para o processo de Análise de Sentimentos, cada etapa realizada é considerada crítica, e deve ser cumprida com o máximo rigor a fim de não degradar o processo. A operação de rotulação dos *tokens* realizada por um *POS Tagger* é a base para à etapa de extração de características. Logo demanda bastante atenção, pois erros na rotulação dos *tokens* acarretam em erros na extração da opinião.

Visão geral de um *POS Tagger*

Um *POS Tagger* é uma ferramenta de processamento de linguagem natural que realiza a etiquetagem dos *tokens* identificados em um corpus. Para se utilizar um *POS Tagger* é necessário que se faça um treinamento da ferramenta a partir de um conjunto de textos pré-rotulados de um dado domínio.

Treinado, o *POS Tagger* é capaz de efetuar a rotulação de termos linguísticos, a partir de textos de diversas origens, para o idioma pelo qual recebeu o treinamento. Porém a linguagem natural é bastante dinâmica e para cada tipo de corpus, existe um padrão de linguagem. Assim, se tem textos mais formais, como por exemplo, os jornalísticos, até textos cujo caráter é mais informal, apresentando gírias e neologismos. Para cada tipo de texto analisado, é necessário que haja um treinamento do *POS Tagger*, a partir de um corpus específico, para que o mesmo realize a sua função com o mínimo de erros possível.

Treinamento realizado e *POS Taggers* considerados para análise com-

parativa

Neste trabalho foi realizado o treinamento de um *POS Tagger* a partir do corpus Amazônia, que é um corpus já rotulado oriundo do sítio colaborativo Overmundo, e está disponível para pesquisadores. O corpus possui aproximadamente 275.000 frases de textos brasileiros, cujo gênero é opinativo, o domínio é cultura brasileira e mescla linguagem formal e informal.

Para treinar o *POS Tagger* foi utilizada uma ferramenta da biblioteca *Apache OpenNLP*. O corpus Amazônia é originalmente encontrado no formato ".ad"(Árvores Deitadas), que o estrutura de uma maneira particular, e para que o treinamento do *POS Tagger* seja efetuado com essa base, primeiro a mesma deve ser convertida para o formato *OpenNLP*. Essa conversão também foi efetuada por uma ferramenta da biblioteca Apache. Após a conversão do formato, foi realizado o treinamento do *POS Tagger* cuja taxa de acerto geral foi de 97,02%.

O segundo *POS Tagger* utilizado na análise, foi o *POS Tagger* padrão da biblioteca *Apache*. A biblioteca fornece inicialmente modelos para o uso de *POS-Taggers* já treinados, para diversos idiomas. Neste projeto o idioma adotado foi o Português. Para esse idioma existe um modelo padrão de *POS Tagger* cujo treinamento foi realizado a partir do corpus Bosque, corpus este que foi desenvolvido pelo mesmo projeto que deu origem ao corpus Amazônia.

O corpus Bosque é composto por 9368 frases cujo gênero do texto é jornalístico, a linguagem utilizada é formal e a sua origem é advinda de textos do Jornal Folha de São Paulo e textos públicos. É totalmente revisado por linguistas, logo esse corpus é bem preciso quanto à rotulação de seus elementos.

E para fechar a comparação da influência do *POS Tagger* no processo de extração de características, foi realizada a rotulação manual das bases de testes utilizadas neste projeto, de modo a simular um *POS Tagger* ideal.

Teste realizado

A partir do protótipo realizado neste trabalho, foi executado o processo de extra-

ção dos pares de opinião (característica,expressão opinativa), em relação às duas bases de testes elaboradas , e para cada base, três execuções foram realizadas: uma utilizando o *POS Tagger* padrão, fornecido pela biblioteca *Apache OpenNLP*, outra utilizando o *POS Tagger* treinado a partir da base Amazônia e por fim, utilizando uma rotulação manual dos dois *corpus* criados neste projeto.

Os resultados dos testes são apresentados na tabela 4.1:

Modelos avaliados	Corpus de Treinamento	Base Produtos		Base Serviços	
		Precisão	Cobertura	Precisão	Cobertura
POS-Tagger Padrão	Bosque	0,349	0,248	0,457	0,218
POS-Tagger Treinado	Amazônia	0,367	0,308	0,438	0,259
Rotulação Manual	-	0,408	0,350	0,533	0,323

Tabela 4.1: Análise comparativa do impacto do *POS Tagger* na acurácia do processo de extração de características

Por essa tabela pode-se perceber que o tipo de corpus utilizado para treinamento do *POS Tagger*, influência diretamente na acurácia do processo de extração de características. O *POS Tagger* padrão, fornecido pela biblioteca Apache, foi o candidato avaliado que apresentou os menores valores de cobertura, e baixos valores de precisão. A explicação para isso está ligada ao corpus em que o *POS Tagger* padrão foi treinado, ou seja, apesar do corpus Bosque ser um corpus revisado, logo com rotulação precisa, o seu tamanho é limitado e o seu gênero é jornalístico, diferente dos textos analisados, que representam mensagens de gênero opinativo. Assim, vários termos não são identificados corretamente no processo de rotulação e como consequência tem-se valores baixos na acurácia.

Para o *POS Tagger* treinado com a base Amazônia, nota-se um pequeno acréscimo nos valores de precisão em relação ao *POS Tagger* anterior quando comparado com os resultados da base de produtos. E um leve decréscimo nos valores de precisão quando comparado com os resultados da base de serviços. Isso ocorre, pois apesar de do corpus Bosque ser um corpus relativamente pequeno, ele é totalmente revisado e bastante preciso em sua rotulação. Porém, a diferença na medida de cobertura é bem significativa. Por ser treinado a partir de um corpus maior, o *POS Tagger* treinado possui uma maior capacidade de classificar corretamente *tokens*, propiciando a extração de mais pares na

etapa de extração de opiniões.

E por fim, a avaliação da rotulação manual foi a mais positiva. Apresentando maiores valores tanto para cobertura, quanto para precisão. O que é esperado, visto que nesse caso o processo não estaria sofrendo perdas por conta de erros no processo de rotulação.

Na tabela 4.2 é mostrado o ganho real nas medidas de precisão e cobertura, para o processo de extração de opiniões, em relação à utilização de um modelo de POS tagger treinado a partir de uma base opinativa e a partir de um POS Tagger ideal (rotulação manual) comparado ao resultado de extração feito a partir do POS Tagger padrão da biblioteca Apache, que não foi treinado a partir de um corpus específico para os domínios dos textos analisados neste projeto.

Modelos avaliados	Corpus de Treinamento	Base Produtos		Base Serviços	
		Precisão	Cobertura	Precisão	Cobertura
POS TaggerTreinado	Amazônia	5,2%	24,2%	-4,2%	18,8%
Rotulação Manual	-	16,9%	41,12%	16,6%	48,16%

Tabela 4.2: Ganho de eficiência por tipo de *POS Tagger* utilizado

Por essa tabela pode-se perceber que o treinamento de um *POS Tagger*, a partir de um corpus específico para o tipo de texto que se pretende analisar, é fundamental. O treinamento do *POS Tagger* realizado neste trabalho, apesar de não ter sido feito a partir de textos opinativos relativos a produtos e serviços, foi realizado a partir de um corpus opinativo de grande dimensão. Desta forma o *POS Tagger* pôde melhorar a sua classificação na rotulação dos *tokens* em relação ao *POS Tagger* Padrão, o que resultou em ganhos tanto nas medidas de precisão quanto de cobertura, com exceção da precisão obtida nos testes realizados sobre a base de serviços utilizando o *POS Tagger* treinado, onde houve um pequeno decréscimo.

Os ganhos são ainda maiores caso se considere um *POS Tagger* ideal (rotulação manual). Melhorando as medidas de cobertura em mais de 40% nos testes realizados a partir das duas bases desenvolvidas.

Isso demonstra que o treinamento adequado de um *POS Tagger* é de grande importância no processo de extração de opiniões e que quanto mais robusta e específica

for à base pelo qual o *POS Tagger* é treinado, menores serão os erros de rotulação, e melhores serão os resultados de acurácia para o processo de extração das opiniões.

4.2.2 Avaliação da influência da abordagem utilizada no processo de extração de Características (opiniões).

Foi visto que a etapa de extração de opiniões é realizada por regras gramaticais, sendo duas técnicas principalmente utilizadas: por proximidade entre substantivos e adjetivos e por padrões fixos.

A técnica por proximidade tem como ponto forte a flexibilidade na extração dos pares. Enquanto a técnica por padrões fixos possui a capacidade de extrair pares de opiniões em formatos mais complexos, permitindo que características e expressões opinativas sejam extraídas na forma de locuções e que expressões opinativas possam ser extraídas em classes gramaticais diversas, não se limitando em serem somente adjetivos.

Para que se pudesse verificar a eficiência de cada técnica para o processo de extração de opiniões, foram criadas três variações do protótipo. A primeira utilizando na etapa de extração de opiniões somente a técnica da proximidade entre substantivos e adjetivos, a segunda utilizando somente a técnica de padrões fixos, considerando os padrões criados neste projeto, e a terceira utilizando a junção das duas técnicas, que representa o protótipo completo desenvolvido neste projeto.

A partir das duas bases de teste criadas, dos domínios de produtos e serviços, realizou-se o processo de extração de opiniões através de cada variação do protótipo criada, cujos resultados de acurácia são apresentados na tabela 4.3:

Abordagens	Base Produtos		Base Serviços	
	Precisão	Cobertura	Precisão	Cobertura
Proximidade	0,313	0,184	0,390	0,165
Padrões Fixos	0,413	0,273	0,529	0,255
Proximidade + Padrões Fixos	0,408	0,350	0,533	0,323

Tabela 4.3: Resultados de acurácia por abordagem

A técnica de Proximidade, para o processo de extração de opiniões, foi a que apresentou os menores valores de precisão e cobertura como resultado final. Mudando

a abordagem e considerando somente a técnica por padrões fixos, nota-se que há uma melhora significativa nas medidas de precisão e cobertura. E por fim, considerando a abordagem adotada neste projeto, que foi utilizar ambas as técnicas anteriores, obtiveram-se os melhores resultados, principalmente com relação à medida de cobertura.

Na tabela 4.4 é feita uma comparação de ganho em eficiência para o processo de extração de opiniões, das abordagens que tiveram melhores valores de acurácia, em relação à abordagem que teve os piores resultados, que foi a que utilizou apenas a técnica de Proximidade.

Abordagens	Base Produtos		Base Serviços	
	Precisão	Cobertura	Precisão	Cobertura
Padrões Fixos	31,9%	51,1%	35,64%	54,54%
Proximidade + Padrões Fixos	30,35%	90,21%	36,66%	95,75%

Tabela 4.4: Ganho em eficiência da abordagem

Utilizando a técnica de Padrões Fixos, a partir dos padrões criados neste projeto, ao invés da técnica da Proximidade, conseguiu-se uma melhora de mais de 50% nas medidas de cobertura e mais de 30% nas medidas de precisão nos testes realizados. Demonstrando que a técnica de Proximidade é limitada e que a técnica de Padrões Fixos, caso seja desenvolvida com um conjunto de padrões adequado, pode ser superior, obtendo melhores valores de acurácia no processo final de extração de opiniões.

A outra abordagem, que considera a utilização de ambas as técnicas, Proximidade e Padrões Fixos, foi ainda mais efetiva em relação à técnica da Proximidade. Promovendo um ganho real de eficiência de mais de 90% nas medidas de cobertura e mais de 30% nas medidas de precisão.

Pelos números, pode-se dizer que a abordagem adotada neste projeto, a união das duas técnicas para o processo de extração de opiniões, é adequada e proporciona maior robustez ao protótipo, fato comprovado pelos melhores resultados de acurácia.

4.3 Considerações Finais

Neste capítulo foram realizados testes que demonstraram a importância do treinamento do *POS Tagger* para os resultados finais de acurácia no processo de extração de

pares de opinião.

Foi avaliada também a eficiência das técnicas para extração das opiniões, onde se comprovou que a utilização das técnicas de Proximidade ou Padrões Fixos isoladamente não representa a melhor abordagem para o processo de extração de opiniões. A utilização das duas técnicas conjuntamente demonstrou ser a abordagem mais adequada, apresentando os maiores valores de acurácia na comparação realizada.

Analisando os números finais de acurácia, pode-se perceber que nos testes realizados tanto os valores de cobertura quanto de precisão estão baixos. Entretanto, isso não é devido à abordagem utilizada no processo de extração de opiniões, e sim a alguns fatores:

- **Erros na rotulação do *POS Tagger*:** para os testes onde foram utilizados o POS Tagger padrão ou o *POS Tagger* treinado, a rotulação não é totalmente correta, assim, no processo de extração de pares de opinião, alguns pares que deveriam ser extraídos não são e ainda há a extração de pares não relacionados ao domínio. Caso os treinamentos do *POS Tagger* fossem realizados a partir de corpus mais específicos e relacionados diretamente ao domínio dos textos que foram analisados a acurácia do processo de extração de opiniões seria maior.

- **Opiniões a serem extraídas em um formato complexo:** há uma série de pares de opiniões a serem extraídos nos corpus de teste que representam um formato complexo, sendo formados por locuções, cujas expressões opinativas representam fatos desejados. Exemplo de pares que deveriam ser extraídos:

  <recurso de acesso pela impressão digital , economiza tempo no acesso>
  <Geral , posso usá-lo por 9 horas sem a necessidade de carregar>

  Esse tipo de opinião só pode ser extraído a partir de padrões fixos. Assim, para que houvesse a extração de maior número de pares de opiniões, deveriam ser criados mais padrões fixos específicos.

- **Falhas em algumas regras:** alguns pares extraídos apesar de serem semanticamente corretos não correspondem ao resultado do *benchmark*. Exemplo:

Saída do algoritmo: <Geral , é o telefone perfeito>
Benchmark: <telefone , perfeito>

Assim, uma revisão de alguns padrões fixos criados é necessária.

- **Falta de padronização (*benchmark* com o algoritmo):** em alguns casos o *benchmark* apresenta pares de opiniões que o algoritmo não seria capaz de extrair. Exemplo:

Saída do algoritmo: <iPhone , muito amigável>
Benchmark:<iPhone 5S , muito amigável>

Como não é realizado o reconhecimento de entidade nomeada, o algoritmo não consegue identificar "iPhone 5S" como uma característica. Assim, uma revisão do *benchmark* se faz necessária, ou uma modificação no próprio protótipo, permitindo que na identificação dos *tokens* seja realizado previamente um processo de reconhecimento de entidade nomeada.

- **Rigor na comparação:** para determinar a acurácia do processo, um algoritmo de comparação foi criado. Para que o par extraído seja considerado correto, ele deve corresponder exatamente ao *benchmark*. Isso faz com que pares que sejam extraídos incompletos, mesmo sendo equivalentes semanticamente ao *benchmark*, sejam considerados errados. Exemplo:

Saída do algoritmo: <tamanho , perfeito>
Benchmark: <tamanho , perfeito para minhas pequenas mãos>

Neste caso, apesar da perda da informação, a polaridade da opinião poderia ser definida corretamente, mas o par é considerado errado pelo algoritmo, reduzindo a acurácia final.

Dada à limitação de tempo, esses fatores considerados não foram revistos e aperfeiçoados, o que provavelmente elevaria substancialmente os resultados de acurácia final. No próximo capítulo, será apresentada a conclusão obtida deste projeto. Onde se indi-

cará as contribuições alcançadas para a área de análise de sentimentos, assim como as dificuldades encontradas nesta pesquisa. E por fim, sugerido possíveis trabalhos futuros.

5 Conclusão

Apesar dos resultados apresentados neste projeto não terem sido tão promissores, isso não invalida a abordagem utilizada para extração de opiniões, que pelos testes realizados demonstrou ser bastante efetiva, tornando o processo de extração de opiniões mais robusto. Na verdade, uma série de fatores contribuiu para que a acurácia final apresentasse valores baixos, principalmente a complexidade dos pares a serem extraídos e o rigor na comparação, onde somente a correspondência total é admitida como correta. Assim, para que a acurácia final fosse mais relevante uma série de revisões deveriam ser realizadas e novos padrões fixos teriam que ser criados. Mas de uma maneira geral esse não era o principal objetivo do projeto e sim a comprovação de que a abordagem utilizada no processo de extração de opiniões fosse efetiva. Fato este que ficou comprovado nos testes realizados.

Como contribuições do trabalho, pode-se dizer que houve uma revisão da literatura, apresentando alguns trabalhos que foram desenvolvidos na área, assim como foi proposta uma abordagem diferente das utilizadas até o momento para o processo de extração de características, propondo a utilização de duas técnicas conjugadas já descritas na literatura, porém utilizadas até então de forma individual. Além da criação de bases de teste para a língua portuguesa, possibilitando que os resultados de novos trabalhos desenvolvidos possam ser comparados com os obtidos neste projeto.

As principais dificuldades para o desenvolvimento deste trabalho se deram principalmente pela falta recursos, como por exemplo, não foram encontrados *benchmarks* relativos a mensagens opinativas em português, logo tiveram que ser desenvolvidas duas bases de testes, consumindo bastante tempo. Também não foram encontrados muitos trabalhos correlatos, principalmente para o idioma português, o que dificultou o estudo inicial do tema abordado. E pode-se dizer de uma maneira geral que o tempo para desenvolvimento do protótipo foi muito curto, mais teste e melhoramentos seriam necessários para que os resultados finais de acurácia fossem mais significativos.

Viu-se que a área de análise de sentimentos é um campo de pesquisa em aberto,

muitos melhoramentos ainda devem ser realizados. Notavelmente é um processo complexo e que demanda por desenvolvimento de novas técnicas e aperfeiçoamento das existentes. Porém, é uma área de pesquisa que promete ser bastante promissora, cujas possibilidades de aplicação do processo ao mundo real são infinitas.

Como sugestão a trabalhos futuros, indica-se o uso da abordagem utilizada neste projeto, somada a alguns melhoramentos como o processo de correção ortográfico e substituição de gírias e abreviações do corpus a ser analisado de forma automática.

Outra melhoria que poderia ser feita, seria adaptar o processo de extração de características para lidar com mensagens comparativas, bem como desenvolver uma técnica mais eficaz para fazer a resolução de correferência, de modo a aumentar a precisão da extração dos pares de opinião.

A parte de mapeamento de características implícitas de forma automática seria outro trabalho interessante de se estudar, visto que a mesma aumenta o grau de detalhamento do processo de análise de sentimentos.

O agrupamento das características extraídas, representando sinônimos, em um único *cluster* é outra questão que não foi realizada neste trabalho e que seria um melhoramento importante em trabalhos futuros, pois com esse processo se teria resultados de sumarização do processo final de análise de sentimentos mais precisos.

E por fim, seria interessante que fosse realizada a integração da análise de sentimentos a motores de busca, de forma a tornar os mecanismos de busca mais interessantes, efetivos e semânticos.

A Apêndice

A.1 Regras Gramaticais

Termos em negrito representam expressões opinativas extraídas e em sublinhado características extraídas.

Essas representam regras básicas. Para cada regra são feitas quatro variações: considerando haver sujeito determinado e afirmação, sujeito determinado e negação, sujeito indeterminado e afirmação e sujeito indeterminado e negação.

Sendo considerado dois casos para sujeito determinado: substantivo ou locução substantiva.

Padrão 100	100	v-fin	**adv**	**adj**	,	**adv**	**adj**	conj-c	**adv**	**adj**
	101	v-fin	**adv**	**adj**	,	**adv**	**adj**	conj-c	**adj**	
	102	v-fin	**adv**	**adj**	,	**adj**	conj-c	**adv**	**adj**	
	103	v-fin	**adv**	**adj**	,	**adj**	conj-c	**adj**		
	104	v-fin	**adv**	**adj**	**prp**	**v-inf**				
	105	v-fin	**adv**	**adj**	conj-c	**adv**	**adj**			
	106	v-fin	**adv**	**adj**	conj-c	**adj**				
	107	v-fin	**adv**	**adj**						
	108	**v-fin**	**adv**	**prp**	**pron-det**	**n**				
	109	**v-fin**	**adv**	**prp**	**n**					
Padrão 200	200	**v-fin**	art	n	prp	n				
	201	**v-fin**	**art**	**n**	**adj**					
	202	**v-fin**	art	n						
	203	**v-fin**	**art**	**pron-det**	**n**					
	204	**v-fin**	**art**	**adj**	**n**					
Padrão 300	300	**v-fin**	**n**	**prp**	**adj**	**n**				
	301	**v-fin**	**n**	**prp**	**art**	**n**				
	302	v-fin	**n**	**adj**	conj-c	**adj**				
	303	v-fin	**n**	**adj**						

Tabela A.1: Regras Gramaticais parte I

Obs: A regra (Padrão 900) é limitada a alguns verbos: “amo”, “recomendo”, “adoro”, “odeio”, “detesto”, “compraria”, “indicaria”.

Padrão 400	400	v-fin	**adj**	,	**adv**	**adj**	conj-c	**adv**	**adj**	
	401	v-fin	**adj**	,	**adv**	**adj**	conj-c	**adj**		
	402	v-fin	**adj**	,	**adj**	conj-c	**adv**	**adj**		
	403	v-fin	**adj**	,	**adj**	conj-c	**adj**			
	404	v-fin	**adj**	conj-c	**adv**	**adj**				
	405	**v-fin**	**adj**	conj-c	**adj**					
	406	v-fin	**v-fin**							
Padrão 500	500	**v-fin**	**v-inf**	**adv**	**adj**					
	501	**v-fin**	**v-inf**	**adj**						
Padrão 600	600	**v-fin**	**adj**	**prp**	**v-inf**	art	n			
	601	**v-fin**	**adj**	**prp**	**v-inf**					
Padrão 700	700	**v-fin**	pron-det	n						
Padrão 800	800	**v-fin**	**prp**	**n**	**adj**					
	801	**v-fin**	**prp**	**n**						
Padrão 900	900	**v-fin**								

Tabela A.2: Regras Gramaticais parte II

símbolo		categoria
n		nome, substantivo
prop		nome próprio
adj		adjectivo
n-adj		flutuação entre substantivo e adjectivo
v	v-fin	verbo finito
	v-inf	infinitivo
	v-pcp	particípio
	v-ger	gerúndio
art		artigo
pron	pron-pers	pronome pessoal
	pron-det	pronome determinativo
	pron-indp	pronome independente (com comportamento semelhante ao nome)
adv		advérbio
num		numeral
prp		preposição
intj		interjeição
conj	conj-s	conjunção subordinativa
	conj-c	conjunção coordenativa

Tabela A.3: Tabela de Símbolos da Gramática

100	**É muito fino, muito frágil e muito caro.**
101	**É muito fino, muito frágil e caro.**
102	**É muito fino, frágil e muito caro.**
103	**É muito fino, frágil e caro.**
104	**É realmente fácil de usar.**
105	**É muito fino e muito frágil.**
106	**É muito fino e frágil.**
107	**É muito fino.**
108	**Funciona bem com meu iPad.**
109	**Cabe facilmente no bolso.**
200	**Adoro** a quantidade de cores oferecidas.
201	**Possui uma interface impressionante.**
202	**Amei** o telefone.
203	**Excedeu as minhas expectativas.**
204	**Proporciona uma ótima nitidez.**

Tabela A.4: Exemplos das Regras Gramaticais I

300	Gera fotos com alta qualidade.
301	Tenho problemas com o desbloqueio.
302	Tem acesso rápido e fácil.
303	Tem acesso rápido.
400	É fino, muito frágil e muito caro.
401	É fino, muito frágil e caro.
402	É fino, frágil e muito caro.
403	É fino, frágil e caro.
404	É fino e muito frágil.
405	É fino e frágil.
406	É fino.
500	Poderia estar mais satisfeito.
501	Poderia ser melhor.
600	Foi fácil de aprender os aplicativos.
601	Foi fácil de aprender.
700	Amo meu telefone.
800	Funciona de modo inconsistente.
801	Espero por mudanças.
900	Recomendo.

Tabela A.5: Exemplos das Regras Gramaticais II

Referências Bibliográficas

Blei, D. M.; NG, A. Y. ; Jordan, M. I. **Latent dirichlet allocation.** 2003.

Cilibrasi, R.; Vitanyi, P. **The google similarity distance.** In: IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, volume VOL. 19, p. 370–383, 2007.

Dicionário do aurélio online. http://www.dicionariodoaurelio.com/, 2008-2014.

Ding, X.; Liu, B. **Resolving object and attribute coreference in opinion mining.** In: In Proceedings of International Conference on Computational Linguistics, 2010.

Hofmann, T. **Probabilistic latent semantic indexing.** In: In Proceedings of Conference on Uncertainty in Artificial Intelligence, 1999.

Hu, M.; Liu, B. **Mining and summarizing customer reviews**. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004.

Jakob, N.; Gurevych, I. **Extracting opinion targets in a single-and cross-domain setting with conditional random fields.** In: In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2010.

Jindal, N.; Liu, B. **Identifying comparative sentenses in text documents**. In: Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval. 2006.

Kobayashi, N.; Inui, K. ; Matsumoto, Y. **Extracting aspect-evaluation and aspect-of relations in opinion mining.** In: In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.

Lafferty, J.; McCallum, A. ; Pereira, F. **Conditional random fields: Probabilistic models for segmenting and labeling sequence data.** In: In Proceedings of International Conference on Machine Learning, 2001.

Lima, D. **Pairextractor: Extração de pares livre de domínio para análise de sentimentos**. 2011.

Liu, B. **Sentiment analysis and subjectivity**. In: Handbook of Natural Language Processing. 2. ed., 2010.

Liu, B. **Sentiment analysis and opinion mining**. In: Sysnthesis Lectures on Human Language Technologies. 2012.

Popescu, A. M.; Etzioni, O. **Extracting product features and opinions from reviews**. In: In Preceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005.

Qiu, G.; Liu, B.; Bu, J. ; Chen, C. **Expanding domain sentiment lexicon through double propagation.** In: In Proceedings of International Joint Conference on Artificial Intelligence, 2009.

Rabiner, L. R. **A tutorial on hidden markov models and select applications in speech recognition.** In: In Proceedings of IEEE, 1989, 1989.

Siqueira, H.; Barros, F. **A feature extraction process for sentiment analysis of opinions on services**. 2011.

Su, Q.; Xu, X.; Guo, H.; Guo, Z.; Wu, X.; Zhang, X.; Swen, B. ; Su, Z. **Hidden sentiment association in chinise web opinion mining.** In: In Proceedings of International Conference on World Wide Web, 2008.

Turney, P. D. **Thumbs up or thumbs down? : semantic orientation applied to unsupervised classification of reviews**. 2002.

Vilicic, F. **O berço do big data**. Editora Abril, 2013.

Wiebe, J. M.; Bruce, R. F. ; Hara, T. P. O. **Development and use of a gold-standard data set for subjectivity classification**. In: In Proceedings of the Association for Computational Linguistics. 1999.