Universidade Federal de Pernambuco

Centro de Informática
Graduação em Ciência da Computação

PairExtractor: Extração de Pares Livre de Domínio para Análise de Sentimentos

Diego Carlos Lucena de Albuquerque Lima

Trabalho de Graduação

Recife, Dezembro de 2011

Universidade Federal de Pernambuco

Centro de Informática
Graduação em Ciência da Computação

PairExtractor: Extração de Pares Livre de Domínio para Análise de Sentimentos

Diego Carlos Lucena de Albuquerque Lima

Trabalho de Graduação apresentado ao Centro de Informática da Universidade Federal de Pernambuco como requisito para obtenção do Grau de Bacharel em Ciência da Computação.

Orientadora: Flavia de Almeida Barros

Recife, Dezembro de 2011

Agradecimentos

Inicialmente, gostaria de agradecer aos meus pais, Renato e Marta, ao meu avô Carlos e a minha irmã Mirella. Devo boa parte do que sou a eles. Obrigado por fazerem parte da minha vida e contribuirem para minha formação como ser humano.

A Prof. Flavia Barros, orientadora deste trabalho, que sempre quando precisei se mostrou atenciosa e prestativa. Este trabalho não teria sido o mesmo sem a sua ajuda.

A Nelson, por toda ajuda com o trabalho, obrigado pelas dicas e material fornecido.

A minha namorada, Tyara, por ser compreensiva nos momentos em que eu estava muito ocupado com projetos da faculdade e não pude lhe dar muita atenção.

Gostaria de agradecer também aos meus colegas de turma, ao qual partilhamos tantos bons momentos durante este curso. Em especial o pessoal da NUTSolutions, Antônio, Bruno, Rafael e Reynaldo. Valeu pessoal!

Por fim, gostaria de agradecer a todos os professores do Centro de Informática, bem como a todos os colaboradores, que foram fundamentais para a minha formação acadêmica. A todos o meu muito obrigado!

Resumo

A quantidade de informação opinativa na internet é cada vez maior. Usuários expressam suas opiniões na internet através de diversos formas como blogs, foruns, redes sociais e sistemas de recomendações.

Estas opiniões são de extrema importância para empresas, que desejam avaliar a aceitação de seus produtos, bem como definir estratégias de mercado para atingir um determinado público-alvo. Porém, a tarefa de coletar, analisar e sumarizar um conjunto de opiniões na Web não é trivial, uma vez que estes dados estão desestruturados e descentralizados.

A Análise de Sentimentos tem por objetivo formalizar o conhecimento obtido com um conjunto de opiniões através do emprego de processamento computacional. Um sistema de Análise de Sentimentos pode ser dividido em quatro etapas: Classificação de subjetividade, extração de características, classificação de sentimentos e sumarização.

Este trabalho tem por objetivo apresentar o *PairExtractor*, um sistema focado na etapa de extração de características, que extrai pares de características e palavras opinativas de forma automática e inovadora.

Um protótipo do *PairExtractor* foi desenvolvido e foram realizados alguns experimentos utilizando diferentes domínios de forma a validar a independência de domínio e as principais técnicas adotadas, com a obtenção de resultados satisfatórios.

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

Abstract

The amount of opinionated information on the Internet is increasing. Users express their opinions on the Internet through many ways such as blogs, forums, social networks and recommendation systems.

These opinions are extremely important for companies who wish to evaluate the acceptance of their products and also to define market strategies to reach a target public. However, the task to collect, analyze and summarize a set of opinions on the Web is not trivial, since these data are unstructured and decentralized.

The Sentiment Analysis aims to formalize the knowledge obtained with a set of opinions by using computational processing. A Sentiment Analysis system can be divided into four steps: Subjectivity classification, feature extraction, sentiment classification and summarization.

This work aims to present the *PairExtractor*, a system focused on the feature extraction step that extracts pairs of features and opinion words in an innovative and automatic way.

A prototype of *PairExtractor* was developed and some experiments were performed using different domains in order to validate the domain-independence and the main techniques that were adopted, with the achievement of satisfactory results.

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

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

A análise das opiniões expressas por um grupo de pessoas sobre um determinado tópico ou objeto sempre foi de grande interesse, tanto para as empresas, que desejam utilizar tal conhecimento para definir aprimoramentos internos e estratégias de mercado, como para os usuários, que utilizam tal informação como base para tomadas de decisão.

Até recentemente, a quantidade de informação disponível para análise era restrita, dependendo dos esforços das empresas em realizar entrevistas, enquetes e grupos focais ou dos usuários em coletar as informações com com parentes e amigos.

Com o advento da internet, a quantidade de informações opinativas disponíveis cresceu consideravelmente, através de uma vasta gama de conteúdo gerado pelo usuário em em blogs, foruns, redes sociais e sistemas de recomendações.

Porém, apesar da Web facilitar o acesso e distribuição de opiniões, a tarefa de identificar, classificar e sumarizar um conjunto de opiniões não é trivial, devido a natureza descentralizada e desestruturada da própria Web. Mesmo os motores de busca disponíveis atualmente, não são suficientes quando se deseja uma síntese de um conjunto de opiniões sobre um determinado tópico ou objeto.

A Análise de Sentimentos (AS) visa usufruir da vasta quantidade de conteúdo disponível na Web, empregando processamento computacional para formalizar o conhecimento obtido de opiniões.

1.1. Contexto e Motivação

Pesquisas com relação a identificação de textos subjetivos já existem desde os anos 80, e alguns trabalhos sobre AS foram publicados nos anos 90. Mas, foi apenas na última década, com a Web 2.0, que houve um grande crescimento no interesse desta área [14] [28] [30].

Atualmente, existe uma vasta gama de conteúdo gerado pelo usuário na Web em diversas fontes. Para uma mesma fonte, podem existir diversos textos opinativos e em diferentes formatos. Além disso, as opiniões são escritas em uma linguagem informal, onde as construções das sentenças podem variar drasticamente dependendo do usuário.

Fica evidente, que apesar do grande volume de informações opinativas disponíveis na Web, a atividade de encontrar fontes relevantes, extrair em sentenças com opiniões, classificá-las e sumarizá-las é uma tarefa árdua para um usuário comum. Para isso, torna-se necessário a concepção e utilização de ferramentas que auxiliem nesta tarefa, tornando o processo menos manual.

Um sistema completo de Análise de Sentimentos pode ser subdividido em quatro etapas: Classificação de subjetividade, extração de características, classificação de sentimentos e sumarização. Atualmente, a etapa de classificação de sentimentos é a mais amplamente estudada [15].

A extração de características, apesar de ser pouco mencionada na maioria dos trabalhos, é uma das etapas mais complexas e importantes da Análise de Sentimentos e pode impactar na qualidade das etapas subsequentes.

1.2. Objetivos e Solução Proposta

Este trabalho tem como objetivo principal a criação de um método de extração de características baseado em técnicas de Inteligência Artificial Simbólica e Processamento de Linguagem Natural.

Apesar das soluções baseadas em aprendizagem de máquina oferecerem, em alguns casos, maior precisão na extração de características, optamos pela abordagem baseada em estatística e linguística, com a finalidade de ganhar independência de domínio da aplicação. Assim, não será necessário a realização de treinamentos para cada novo domínio de aplicação.

O método aqui proposto teve como ponto de partida o trabalho apresentado por Siqueira [26], que descreve o *WhatMatter*, um sistema para extração de características de um dado domínio a partir de opiniões sobre serviços. O método descrito neste trabalho estende o trabalho de Siquiera, retornando pares na forma (palavras opinativas, característica), a fim de melhorar a precisão da etapa de classificação de sentimentos.

A necessidade de se trabalhar com pares se deve ao fato de que uma mesma palavra opinativa pode ser interpretada de maneira oposta, dependendo do domínio da aplicação (por exemplo, quente é positivo quando se trata de pizza, e negativo quando se trata de cerveja). Assim, a abordagem aqui proposta contribuirá para melhorar os resultados obtidos nas etapas seguintes em um sistema de Análise de Sentimentos (i.e., classificação de sentimentos e sumarização).

A solução proposta neste trabalho foi implementada, resultando em um protótipo do sistema *PairExtractor*. Ressaltamos que não foi encontrado na literatura disponível nenhum trabalho de extração de características semelhante ao aqui apresentado.

1.3. Organização do Trabalho

Este trabalho está dividido em 5 capítulos, da seguinte forma:

O capítulo 1 consiste na parte introdutória do trabalho, onde são apresentados o contexto da Análise de Sentimentos e a inserção da solução proposta neste contexto.

O capítulo 2 apresenta brevemente os principais conceitos da AS, as suas etapas e aplicações. Também é efetuada uma revisão literária de algumas técnicas aplicadas a cada etapa, focando principalmente na etapa de extração de características.

O capítulo 3 descreve a solução proposta, o *PairExtractor*. Neste capítulo é apresentado os módulos que compoem a solução, descrevendo as técnicas e procedimentos adotados para cada um destes módulos.

O capítulo 4 apresenta os experimentos realizados de forma a avaliar o *PairExtractor*. Também são apresentados os resultados obtidos e a análise crítica da solução.

O capítulo 5 contém a conclusão do presente trabalho, apresentando as contribuições deste trabalho para a AS, as dificuldades encontradas e os possíveis trabalhos futuros.

2. Análise de Sentimentos

A Análise de Sentimentos (AS), também denominada de mineração de opiniões, consiste no estudo computacional de opiniões, sentimentos e emoções expressas em texto [15].

Um sentimento pode ser visto como valores pessoais (positivos, negativos ou neutros) expressos sobre uma determinada entidade. Porém, um sentimento normalmente não é expresso de forma clara e precisa. A Análise de Sentimentos basicamente tenta inferir os sentimentos de pessoas baseado nas expressões de linguagem [11].

Neste capítulo, será efetuado um levantamento teórico da Análise de Sentimentos. Na Seção 2.1 serão mostrados alguns dos conceitos básicos, definições e terminologias utilizadas neste trabalho. A Seção 2.2 apresenta os níveis de tratamento da AS. As Seções 2.3 e 2.4 apresentam as etapas da AS. Na Seção 2.5 são descritas algumas aplicações práticas para os conceitos da AS. A Seção 2.6 contém uma breve descrição dos desafios e limitações da área atualmente. Por fim, a Seção 2.7 contém as considerações finais deste capítulo.

2.1. Conceitos Básicos

Nesta seção, serão descritos alguns dos principais conceitos e definições que envolvem a Análise de Sentimentos, necessários para um bom entendimento deste trabalho. Uma vez que a AS possui uma variação de terminologia utilizada na literatura, também será definida a terminologia utilizada neste trabalho.

O Exemplo 2.1 ilustra um texto típico contido em sites de *reviews* sobre produtos, onde pode-se introduzir o problema da Análise de Sentimentos e aplicar alguns dos conceitos da área.

"(1) Eu comprei uma câmera fotográfica há alguns dias. (2) Eu gostei bastante. (3) A qualidade da foto é muito boa. (4) Os vídeos também são nítidos. (5) Porém, a bateria dura pouco. (6) Minha mãe achou a câmera fotográfica muito cara."

Exemplo 2.1 *Review* sobre uma câmera fotográfica

Podemos observar no Exemplo 2.1, composto por 6 sentenças, que a sentença 1 não contém opinião, ou seja, consiste em uma sentença objetiva, composta apenas por fatos. As demais sentenças são subjetivas, contendo opiniões positivas (2, 3 e 4) ou opiniões negativas (5 e 6).

Em seguida, podemos notar que todas as opiniões são expressadas sobre um objeto de forma explícita ou implícita. O objeto da sentença 2 é a câmera fotográfica como um todo e está de forma implícita. Nas sentenças 3, 4 e 5 os objetos são "qualidade da foto", "vídeo" e "bateria", respectivamente. Na

sentença 6 o objeto é preço da máquina fotográfica, que está de forma implícita.

Por fim, podemos verificar que toda opinião possui um portador de opinião, ou seja, a pessoa que emitiu a opinião. Nas sentenças 2, 3, 4 e 5, o portador de opinião é o próprio autor da *review*. Na sentença 6, o portador de opinião é a mãe do autor.

Após a exemplificação do problema da Análise de Sentimentos, podemos definir formalmente os principais conceitos que envolvem este domínio.

2.1.1. Objeto

Um *objeto o* é uma entidade que pode ser um produto, pessoa, evento, organização ou tópico. É associado com um par *o:(C,A)*, onde *C* é uma hierarquia de *componentes* (ou *partes*) e *sub-componentes*, e *A* é um conjunto de *atributos* de *o*. Cada *componente c* tem seu proprio conjunto de *sub-componentes* e *atributos* [15].

Um objeto pode ser representado como uma árvore, como ilustra a Figura 2.1. A raiz da árvore consiste no próprio objeto, e os demais nós podem ser componentes, sub-componentes e atributos deste objeto. Uma opinião pode ser expressa sobre qualquer nó da árvore.

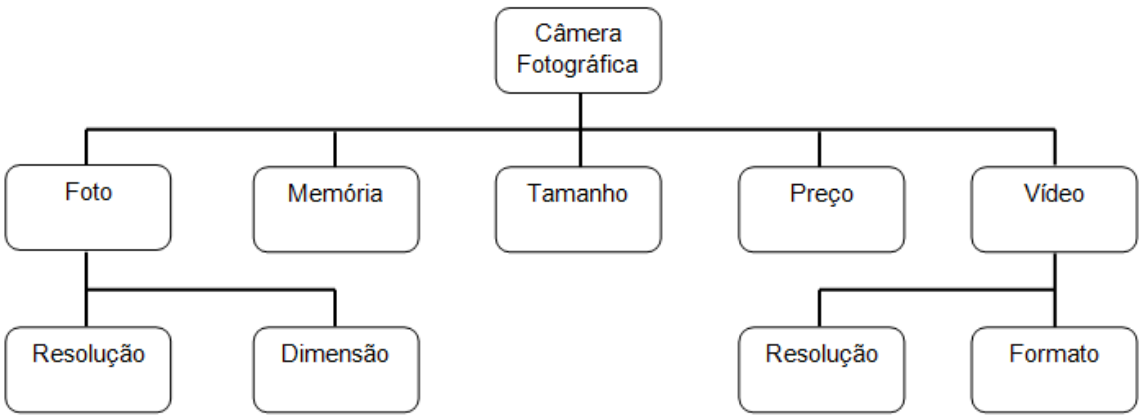

Figura 2.1 Árvore de características

Fonte: Elaboração Própria

De forma a simplificar a definição, neste trabalho tanto os *componentes* quanto os *atributos* serão denominados de *características*.

2.1.2. Característica

Uma *característica* é um componente, sub-componente, atributo, propriedade, parte ou seção de um *objeto*, seja este um produto, pessoa, evento, organização ou tópico [15].

Desta forma, um *objeto o* contém um conjunto finito de *características* $C=\{c_1, c_2, ..., c_n\}$, que inclui o próprio objeto como uma característica especial. Cada característica $c_i \in C$ pode ser expressa com um com uma das palavras de um conjunto finito de *sinônimos* $S_i = \{S_{i1}, S_{i2}, ..., S_{im}\}$ ou indicada por um conjunto finito de *indicadores* $I_i = \{I_{i1}, I_{i2}, ..., I_{iq}\}$.

Uma característica pode ser classificada como *explícita* ou *implícita*. Uma *característica c* é dita *explícita* em uma *sentença s* se *c* ou algum dos seus *sinônimos* do conjunto finito *S* aparecem em *s*. Uma *característica c* é dita implícita em uma *sentença s* se nem *c* e nenhum dos seus *sinônimos S* aparecem em *s*, mas *c* está implícita em *o* e pode ser mapeada através de um dos *indicadores* do conjunto finito *I*.

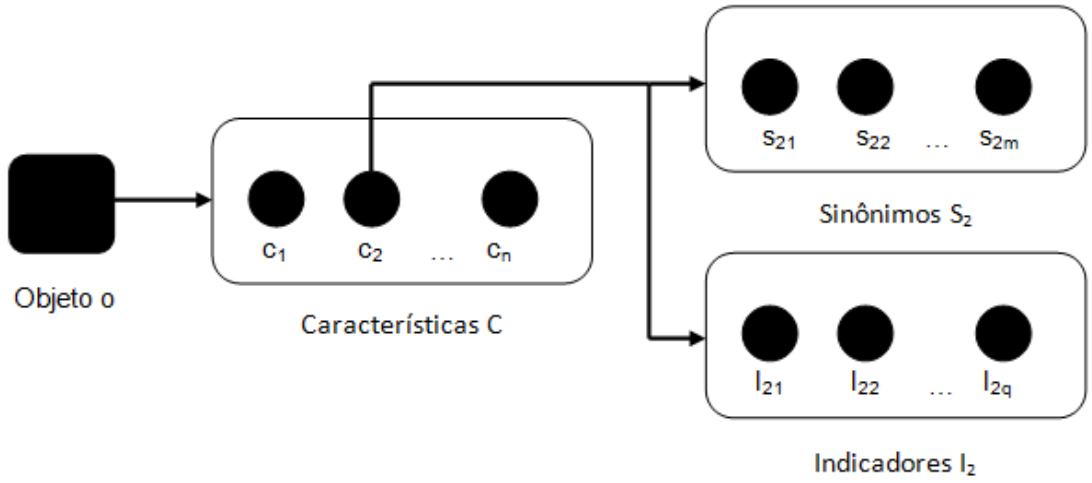

Figura 2.2 Modelo de características, sinônimos e indicadores

Fonte: Elaboração Própria

2.1.3. Portador de Opinião

Um *portador de opinião* é a pessoa ou organização que expressa a opinião [15].

2.1.4. Polaridade

A *polaridade* de uma opinião sobre uma *característica c* indica se a opinião é positiva, negativa ou neutra [15].

2.1.5. Opinião

Antes de definir o conceito de opinião, é importante ter em mente os conceitos de sentença objetiva, subjetiva e opinativa.

Uma *sentença objetiva* expressa informações factuais, ou seja é composta apenas por fatos. O Exemplo 2.2 ilustra uma sentença objetiva.

"Eu comprei uma câmera fotográfica."

Exemplo 2.2 Sentença objetiva

Uma *sentença subjetiva* descreve sentimentos, avaliações, crenças e experiências de pessoas sobre entidades, eventos e suas propriedades. O Exemplo 2.3 ilustra uma sentença subjetiva.

"Eu acho que ele comprou uma câmera fotográfica."

Exemplo 2.3 Sentença subjetiva

Uma *sentença opinativa* é uma sentença que expressa uma opinião positiva ou negativa, de forma explícita ou implícita. Pode ser uma sentença *objetiva* ou *subjetiva*. O Exemplo 2.4 ilustra uma sentença opinativa.

"A câmera fotográfica quebrou em uma semana."

Exemplo 2.4 Sentença opinativa

Apesar de que na maioria dos casos as sentenças opinativas são sentenças subjetivas, podemos ter algumas exceções. Podemos notar no Exemplo 2.3, que apesar da sentença ser subjetiva, ela não expressa nenhuma opinião e que apesar da sentença no Exemplo 2.4 conter apenas fatos, ela implicitamente indica uma opinião negativa sobre a câmera fotográfica.

Com base nas definições precedentes, podemos definir que uma *opinião o* sobre uma *característica c*, consiste em uma visão, atitude, emoção ou avaliação *positiva*, *negativa* ou *neutra* de *c*, de forma *subjetiva* ou *objetiva*, *explícita* ou *implícita,* por um *portador de opinião p* [15].

2.1.6. Palavra Opinativa

Palavras opinativas são palavras que normalmente são utilizadas para expressar sentimentos positivos ou negativos. Geralmente as palavras opinativas são adjetivos ou advérbios, mas também podem ser substantivos ou verbos [15].

2.2. Níveis da Análise de Sentimentos

A Análise de Sentimentos pode ser tratada em níveis distintos e para cada nível podemos ter objetivos diferentes [16].

2.2.1. Nível de Documento

Neste nível, o documento por inteiro é considerado como uma unidade básica de informação. Desta forma, a tarefa principal consiste na classificação global de sentimentos como positiva ou negativa (para o nível de documento, a classificação neutra é normalmente descartada).

Para a classificação de sentimentos neste nível, algumas pressuposições são feitas: É assumido que o documento é opinativo, que foca em apenas um único objeto e que contém a opinião de apenas um único portador de opinião.

Portanto, o nível de documento não é apropriado para qualquer tipo de aplicação, sendo mais explorado na literatura para classificação de sentimentos em sites de *reviews*, que geralmente satisfazem essas pressuposições.

2.2.2. Nível de Sentença

Como visto anteriormente, para algumas aplicações, o nível de documento é muito pouco detalhado, sendo mais apropriado o nível de sentença, que considera cada sentença como unidade básica de informação.

A Análise de Sentimentos no nível de sentença é realizada basicamente em duas etapas: Classificação de subjetividade e classificação de sentimentos. Na classificação de subjetividade são identificadas quais sentenças são subjetivas e estas serão utilizadas no passo seguinte, onde serão classificadas como positiva, negativa ou neutra.

Neste nível, também são feitas pressuposições, assumindo que uma sentença só possui uma opinião, o que é verdade para a maioria das sentenças simples, porém geralmente não se verifica em sentenças compostas e complexas.

2.2.3. Nível de Característica

Um documento ou sentença contendo uma classificação positiva sobre um objeto não significa que o portador de opinião possui opiniões positivas para todas as características deste objeto. Para este tipo de análise, deve-se ir mais fundo, no nível de característica.

Neste nível, a característica é considerada como unidade básica de informação. Além das etapas de classificação de subjetividade e classificação de sentimentos, é introduzida uma nova etapa, que consiste na identificação e extração das características.

Apesar deste nível ser o mais completo, também é o mais complexo, devido ao maior número de etapas e nível de detalhe.

2.3. Etapas da Análise de Sentimentos

Nesta seção serão apresentadas em detalhes, as etapas que compõem um sistema completo de Análise de Sentimentos. Estas etapas normalmente são executadas de forma sequencial, mas em algumas abordagens na literatura [16] [20], estas etapas podem ser combinadas e executadas simultaneamente.

O Exemplo 2.5 será utilizado para ilustrar o funcionamento de cada etapa da Análise de Sentimentos.

> *"Minha mãe deu uma câmera fotográfica de presente. A qualidade da foto não é ótima, mas a duração da bateria é longa. A memória é boa e a câmera fotográfica não é grande."*

Exemplo 2.5 *Review* de entrada para um sistema de Análise de Sentimentos

2.3.1. Classificação de Subjetividade

Em algumas aplicações envolvento a Análise de Sentimentos, um texto recebido como entrada é assumido como sendo um texto opinativo. Porém em diversas outras aplicações, é necessário decidir onde há ou não subjetividade, indicando a provável presença de opiniões.

> *"Minha mãe deu uma câmera fotográfica de presente.* ***A qualidade da foto não é ótima, mas a duração da bateria é longa. A memória é boa e a câmera fotográfica não é grande."***

Exemplo 2.6 Sentenças subjetivas selecionadas na classificação de subjetividade

No Exemplo 2.6, podemos notar que apenas 2 das sentenças foram selecionadas como sentenças opinativas, através da classificação de subjetividade. Na maioria dos trabalhos, as sentenças objetivas são descartadas nesta etapa.

Muitos trabalhos, focam na resolução deste problema, que pode ter um impacto direto na qualidade das etapas subsequentes.

Hatzivassiloglou e Wiebe [13] propuseram uma solução de identificação de subjetividade baseado na orientação dos adjetivos contidos em uma sentença.

Yu e Hatzivassiloglou [36] conseguiram um percentual de acerto de 97% com um método de classificação Bayesiana em um *corpus* consistindo de artigos do Wall Street Journal.

Junior [10] realizou estudos comparativos de abordagens baseadas em aprendizagem de máquina e de abordagens baseadas em estatística e linguística para a classificação de subjetividade. Também foi proposto um

método para a classificação de subjetividade baseado em estatística e linguística.

2.3.2. Extração de Características

Após a identificação de sentenças opinativas através da classificação de subjetividade, a etapa seguinte consiste na extração das características. Esta etapa é importante para identificar quais são as características referenciadas por um portador de opinião e através destas características, realizar uma análise com maior nível de detalhe nas etapas seguintes.

*".. A qualidade da **foto** não é ótima, mas a duração da **bateria** é longa. A **memória** é boa e a câmera fotográfica não é grande **(tamanho)**."*

Exemplo 2.7 Características selecionadas na extração de características

Através do Exemplo 2.7, podemos perceber quais são as características avaliadas pelo portador de opinião. Podemos perceber também, que existem características que estão implícitas (no caso, tamanho) e devem ser referenciadas através de um indicador de característica. Pouco trabalho tem sido feito para mapear características implícitas [16].

Devido ao fato da precisão da etapa de extração de características estar associada ao contexto da opinião e a semântica das palavras, a etapa de extração de características pode ser considerada como uma das mais complexas na Análise de Sentimentos [26].

Na Seção 2.4 algumas abordagens para a extração de características serão apresentadas com maior detalhe.

2.3.3. Classificação de Sentimentos

Com as características avaliadas pelo portador de opinião devidamente extraídas, o próximo passo consiste classificar cada uma dessas características através da determinação de suas polaridades.

Uma das abordagens mais simples e largamente utilizada na literatura, consiste na abordagem baseada em *lexicon* [4], onde são aplicadas um conjunto de regras para determinar a polaridade de cada palavra opinativa. Esta abordagem funciona da seguinte forma:

- Para cada palavra opinativa positiva, é associado a polaridade +1, para cada palavra opinativa negativa, é associado a polaridade -1 e, para cada palavra opinativa dependente de contexto, é associado a polaridade 0.
- Se houverem palavras de negação próximas às palavras opinativas, então a polaridade é invertida.

- Se houverem palavras adversativas como “mas”, “no entanto”, etc., é aplicada a regra na qual a frase anterior e a frase posterior às palavras adversativas devem possuir polaridades opostas.
- Se houverem palavras aditivas como “e”, “como também”, etc., é aplicada a regra na qual a frase anterior e a frase posterior às palavras aditivas devem possuir a mesma polaridade.
- Por fim, é aplicada uma função de agregação para computar a orientação final da opinião.

*“.. A qualidade da foto não é ótima**[+1]**, mas duração da bateria é longa**[0]**. A memória é boa**[+1]** e a câmera fotográfica não é grande**[0]**.”*

*“.. A qualidade da foto não é ótima**[-1]**, mas duração da bateria é longa**[0]**. A memória é boa**[+1]** e a câmera fotográfica não é grande**[0]**.”*

*“.. A qualidade da foto não é ótima**[-1]**, mas duração da bateria é longa**[+1]**. A memória é boa**[+1]** e a câmera fotográfica não é grande**[0]**.”*

*“.. A qualidade da foto não é ótima**[-1]**, mas duração da bateria é longa**[+1]**. A memória é boa**[+1]** e a câmera fotográfica não é grande**[+1]**.”*

Exemplo 2.8 Abordagem baseada em *lexicon*

O Exemplo 2.8 ilustra a polaridade aplicada a cada palavra opinativa durante cada etapa da classificação de sentimentos.

Silva [25] propõe um método para a classificação de sentimentos de forma a automatizar a tarefa de atribuição da polaridade das palavras opinativas.

2.3.4. Sumarização

A última etapa de um sistema completo de Análise de Sentimentos consiste na sumarização dos dados. Esta etapa é de grande importância, pois o modo como as informações serão exibidas pode comprometer a eficiência das etapas anteriores e não auxiliar devidamente o usuário na tomada de decisão.

A sumarização pode ser efetuada basicamente de duas maneiras: de forma textual ou de forma gráfica. A sumarização textual é a abordagem que tem sido mais utilizada na literatura e normalmente consiste em adaptar técnicas já existentes de sumarização de documentos para a Análise de Sentimentos [20]. A Figura 2.3 ilustra um exemplo de sumarização realizado de forma textual.

Bottom line, well made camera, easy to use, very flexible and powerful features to include the ability to use external flash and lense/filters choices. It has a beautiful design, lots of features, very easy to use, very configurable and customizable, and the battery duration is amazing! Great colors, pictures, and white balance. The camera is a dream to operate in automode, but also gives tremendous flexibility in aperture priority, shutter priority, and manual modes. I'd highly recommend this camera for anyone who is looking for excellent quality pictures and a combination of ease of use and the flexibility to get advanced with many options to adjust if you like.

Figura 2.3 Sumário textual gerado automaticamente

Fonte: [20]

A sumarização textual nem sempre é a solução mais intuitiva, pois uma grande quantidade de texto pode deixar o usuário final confuso, principalmente quando deseja-se analisar os dados de forma estatística. Neste caso, uma sumarização gráfica é mais adequada. A Figura 2.4 ilustra um exemplo de sumarização gráfica.

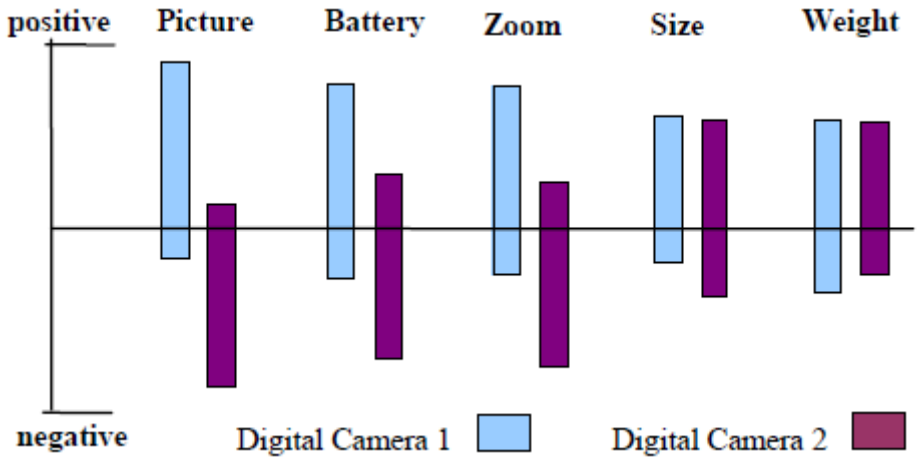

Figura 2.4 Sumário gráfico comparativo

Fonte: [14]

Alguns trabalhos focam na sumarização através da representação gráfica de dados estatísticos obtidos através da Análise de Sentimentos como é o caso do *Opinion Observer* [14], que permite uma rápida visualização comparativa das diversas características de objetos de marcas diferentes.

2.4. Extração de Características

Nesta seção, a etapa de extração de características será mostrada em maior detalhes, que é o foco principal da solução proposta neste trabalho, o *PairExtractor*.

As pesquisas atuais na extração de características são realizadas principalmente no área de *reviews* de produtos online [15]. As técnicas empregadas para a extração de características varia de acordo com o formato das *reviews*. Basicamente, as *reviews* podem ser encontradas em 3 formatos:

1. **Prós e contras**: Neste formato o usuário é solicitado a descrever os prós e contras de um produto separadamente.
2. **Prós, contras e *review* detalhado:** Além de descrever os prós e contras separadapente, o usuário é solicitado a descrever uma *review* detalhada.
3. **Formato livre:** Neste formato, o usuário tem liberdade de escrever sua *review* livremente.

As *reviews* do tipo 1 tendem conter textos breves, enquanto as *reviews* do tipo 3 geralmente contém sentenças completas e elaboradas. O tipo 2, é um meio termo, onde podem ser encontrados tanto textos breves, quanto textos mais elaborados.

Para o formato 1, geralmente são empregadas técnicas baseadas em aprendizagem de máquina, já para o formato 3, geralmente são empregadas técnicas baseadas em estatística e linguística. Para o formato 2, a técnica empregada dependerá do contexto da aplicação.

A seguir, serão apresentadas algumas abordagens para a extração de características.

2.4.1. Abordagens baseadas em aprendizagem de máquina

Qiu [22] apresenta um método baseado em aprendizagem de máquina, denominado *Double Propagation*.

A idéia central do *Double Propagation* consiste em explorar relações sintáticas entre palavras opinativas e características. Foi mostrado que palavras opinativas quase sempre estão associadas às características de alguma maneira.

Para o funcionamento do método, são utilizadas um conjunto de palavras opinativas como semente e através das palavras opinativas, novas características são identificadas e vice-versa.

Outra técnica, descrita por Liu [14], consiste na utilização de regras, que são geradas através de padrões sequenciais na Análise de Sentimentos.

Uma regra pode ser definida da seguinte forma: $X \rightarrow Y$, onde Y é uma sequência e X é uma sequencia produzida por Y, substituindo-se alguns dos seus itens por curingas, que podem ser qualquer item.

O processo consiste em converter cada sentença em uma sequência de elementos, onde cada elemento é composto por uma palavra e uma tag, que indica sua classe gramatical. Em seguida, as características são extraídas através do casamento de padrão com as regras que foram previamente geradas.

Apesar de geralmente apresentarem bons resultados, as técnicas que envolvem a utilização de aprendizagem de máquina, tem a grande desvantagem de necessitar de uma grande base para realizar o treinamento.

2.4.2. Abordagens baseadas em estatística e linguística

Uma abordagem baseada em frequência de termos candidatos a características é descrita por Hu [7]. O autor utiliza como base o fato de que substantivos frequentemente mencionados pelos portadores de opinião são normalmente características pertinentes a um determinado objeto.

Após as características frequentes serem identificadas, o passo seguinte consiste em identificar características infrequentes, que podem ser úteis para determinadas aplicações. Para isto, são utilizadas as palavras opinativas próximas às características frequentes, de forma a encontrar características infrequentes que também estejam próximas destas palavras opinativas.

Esta abordagem possui o incoveniente de necessitar de um corpus grande para que o resultado seja satisfatório. Pois, com uma quantidade pequena de opiniões, o limiar entre uma característica frequente e uma característica infrequente é muito tênue.

Os trabalhos descritos por Siqueira [26] e Popescu [21], visam aprimorar a abordagem previamente descrita através da utilização de uma medida denominada PMI (*Pointwise Mutual Information*) [34].

Através desta medida, é possível verificar o quão relacionadas são duas palavras de forma a filtrar alguns substantivos que não estão associados ao domínio e provavelmente não são características verdadeiras. Esta técnica é descrita com maior detalhes no Capítulo 4, onde são realizados alguns experimentos com o objetivo de avaliar seu desempenho

2.5. Aplicações

São inúmeras as aplicações possíveis na área de Análise de Sentimentos, fato este que ajudou a impulsionar as pesquisas nesta área.

Esta seção tem por objetivo ilustrar algumas das possíveis aplicações dos conceitos da AS na prática.

2.5.1. Sites de *Reviews*

A Análise de Sentimentos pode ser utilizada como alternativa a sites de *reviews* que solicitam *feedback* sobre o que os usuários pensam sobre determinado produto, como é o caso do site Epinions.com. Este processo de coleta de opiniões poderia ser efetuado de forma proativa e automática.

Pode-se ainda citar outra aplicação da AS para aperfeiçoamento em sites de *Reviews*, que consiste na correção automatica de *ratings*. Caso um usuário acidentalmente submetesse uma avaliação baixa de um determinado produto, poderia ser verificado que o texto contém uma avaliação positiva e o *rating* seria automaticamente corrigido.

2.5.2. Inteligência de *Marketing*

Esta é uma das aplicações que despertam um maior interesse pelas corporações. Através da AS, as empresas podem coletar opiniões de diversos tipos de sites na Web, como blogs e sites de *reviews*, gerando relatórios automaticamente, reduzindo consideravelmente o tempo de análise das informações e permitindo a definição de estratégias de *marketing* para determinado produto em um prazo menor.

2.5.3. Política

A opinião coletiva é um assunto de grande interesse na política. A AS pode ser utilizada para coletar a opinião de um público sobre determinado político ou ação política.

Outra aplicabilidade da AS neste domínio, consiste na predição da afiliação política, ou seja, determinar o apoio ou rejeição das figuras públicas a respeito de propostas políticas com base em informações extraídas em listas de discussão de caráter político [6].

2.6. Desafios e Limitações

Por ser uma área de estudo relativamente nova, a Análise de Sentimentos possui muitos problemas ainda não tratados. Além disto, por consistir em um subproblema do Processamento de Linguagem Natural, a AS herda parte dos desafios implicados nesta área .

Muitos dos desafios e limitações da AS transcendem questões tecnológicas e são inerentes a questões lingüísticas e culturais. As sutilezas de linguagem são os principais fatores de complexidade na AS.

A utilização de diversos recursos de linguagem como a ambiguidade, negação, ironia e sarcasmo, tornam tênues os limites entre fatos e opiniões.

Também existem problemas relativos a atribuição de significados às palavras, como a sinonímia, onde um mesmo significado pode estar aplicado a palavras diferentes e a polissemia, onde vários significados são aplicados a uma mesma palavra.

Outro problema, consiste na utilização de co-referência através da utilização de pronomes para referênciar substantivos mencionados anteriormente no texto.

Pode-se ainda citar um problema inerente a internet, onde os textos geralmente são escritos de forma informal, muitas vezes utilizando um vocabulário específico da Web composto por diversas gírias e neologismos.

Outra limitação da área é a capacidade de aperfeiçoamento das técnicas propostas. Estudos têm mostrado que a precisão e qualidade dos resultados obtidos com Processamento de Linguagem Natural não crescem na mesma proporção do esforço empregado para que isso aconteça [12].

2.7. Considerações Finais

Neste capítulo foram mostrados os principais conceitos envolvendo a Análise de Sentimentos, onde este processo foi dividido em quatro etapas: Classificação de subjetividade, extração de características, classificação de sentimentos e sumarização.

A etapa de extração de características foi vista com maior nível de detalhes, devido ao fato de ser o foco principal da solução proposta. Foram mostradas algumas técnicas utilizadas na literatura.

No capítulo a seguir, será descrito em detalhes a solução proposta neste trabalho para extração de pares, o *PairExtractor*.

3. PairExtractor

Nos capítulos precedentes, foi visto que a demanda por aplicações na área de Análise de Sentimentos é cada vez maior e que a etapa de extração de características, apesar de pouco mencionada na literatura em comparação com as etapas de classificação de subjetividade e de sentimentos, tem grande importância no resultado final de um sistema completo de Análise de Sentimentos.

Neste capítulo, será descrito em detalhes a solução proposta para a extração de pares, o *PairExtractor*. Na Seção 3.1 será descrito o problema ao qual o *PairExtractor* se propõe a solucionar. A Seção 3.2 descreve a abordagem de extração de pares adotada. Na Seção 3.3 é apresentada a arquitetura geral do sistema. Da Seção 3.4 à 3.7 serão mostrados os módulos e etapas que compõem o *PairExtractor*. Por fim, na Seção 3.8 serão descritas as considerações finais deste capítulo.

3.1. Caracterização do Problema

Foi visto anteriormente que um sistema de Análise de Sentimentos é um sistema complexo, composto por até quatro etapas. Desta forma, por questões de restrição de tempo, o escopo do sistema descrito neste trabalho limita-se à etapa de extração de características, que consiste em uma das etapas mais importantes da AS.

A solução apresentada por Siqueira [26] descreve um método de extração de características de forma isolada ao domínio da aplicação. Isto quer dizer que o contexto ao qual as características estão situadas não é levado em consideração nas etapas seguintes da Análise de Sentimentos. O Exemplo 3.1 ilustra um tipo de situação em que isto ocorre.

*"Eu achei a cerveja **gelada**" (*Positivo)
"*A pizza estava **gelada**."* (Negativo)

Exemplo 3.1 Polaridade de palavras opinativas sensíveis ao contexto

Podemos observar que para uma mesma palavra opinativa "gelada", pode-se ter duas avaliações de polaridade diferentes, dependendo do contexto indicado pela característica (cerveja ou pizza).

Desta forma, o ideal seria que fossem extraídos não apenas características, mas também as palavras opinativas associadas a cada característica na forma de pares (características, palavras opinativas) com o intuito de aprimorar a classificação de sentimentos através do cálculo de polaridade.

O *PairExtractor* foi projetado com a finalidade de receber um corpus de opiniões como entrada, realizar o processo de extração de características e palavras opinativas associadas a cada característica, resultando como saída em um conjunto de pares na forma de (característica, palavra opinativa). Na literatura relacionada, não foi encontrado nenhum outro método de extração de características que fosse concebido desta forma.

O idioma selecionado para o *PairExtractor* foi o inglês, devido ao fato de ser o idioma mais utilizado no mundo acadêmico, e por possuir um maior volume de dados disponíveis para elaborações de *corpus*. Além disso, podemos contar com uma grande variedade de ferramentas de processamento de linguagem natural para essa língua, como bases léxicas, *POS-taggers* e *parsers*.

A solução foi concebida de forma a facilitar uma possível integração com outras técnicas e etapas, com o propósito de constituir um sistema completo de Análise de Sentimentos.

3.2. Abordagem Adotada

A abordagem adotada para o *PairExtractor*, consiste na abordagem baseada em estatística e linguística, descrita nos capítulos anteriores.

Pelo fato do *PairExtractor* ser uma solução independente de domínio para textos em formato livre, esta abordagem se mostrou mais apropriada, pois no caso de uma abordagem baseada em aprendizagem de máquina, seria muito custoso etiquetar uma base de treinamento em cada domínio ao qual o sistema fosse utilizado.

3.3. Arquitetura Geral

A Figura 3.1 ilustra a arquitetura geral do *PairExtractor*. O sistema foi projetado de forma modular, visando uma melhor manutenbilidade, legibilidade e extensibilidade. Os módulos que compõem o sistema são: Pré-processamento, extração de pares e filtro de pares.

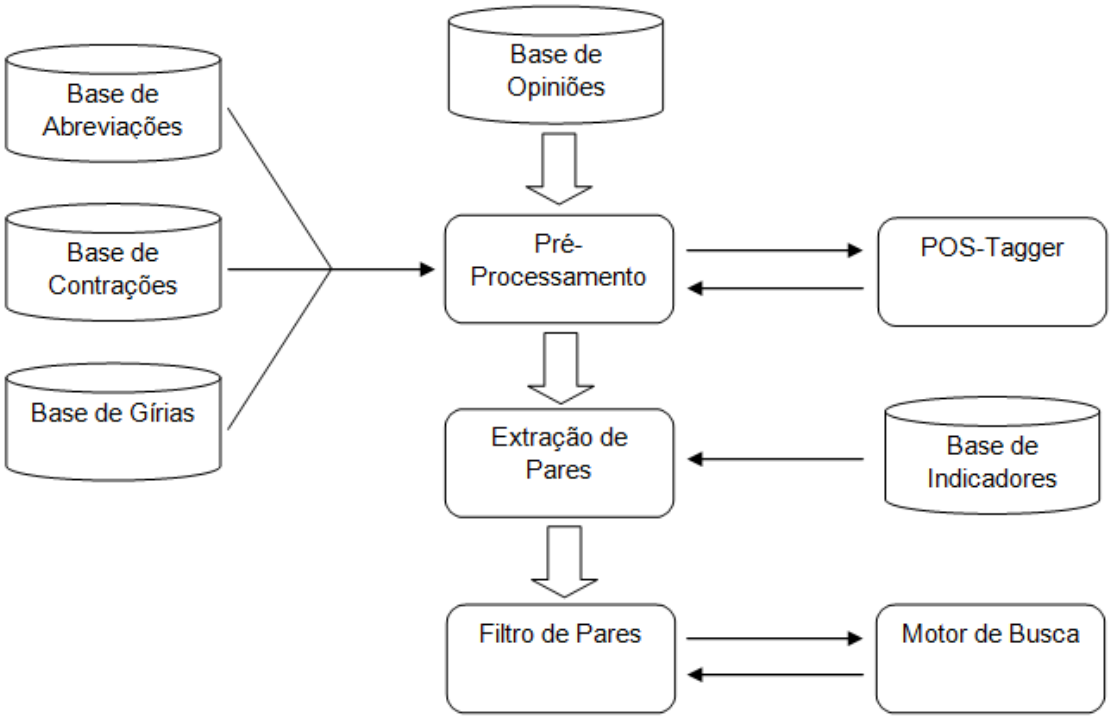

Figura 3.1 Arquitetura geral do *PairExtractor*

Fonte: Elaboração Própria

O módulo de pré-processamento realiza as operações necessárias para que a entrada fornecida esteja no formato adequado e devidamente etiquetado para a posterior utilização dos dados pelo módulo de extração de pares.

O segundo módulo é responsável pela extração de pares propriamente dito, através da identificação dos substantivos frequentes e de alguns infrequentes com base nas palavras opinativas próximas aos substantivos frequentes.

O último módulo consiste no filtro de pares, que realiza uma otimização no processo de extração de pares, de forma a eliminar pares com características possivelmente falsas.

Além dos módulos principais, também são utilizadas diversas bases de dados pelo sistema e algumas ferramentas externas, que serão descritas com maior detalhe nas seções que seguem.

3.4. Base de Opiniões

A base de opiniões contém as opiniões que serão utilizadas no processo de extração de pares. Estas opiniões podem ser extraídas de forma manual ou por algum Web *crawler*, que consiste em um programa que varre um conjunto de páginas da Web e exrtrai informações relevantes de forma automática.

O sistema utiliza como entrada um arquivo com um conjunto de opiniões. Para este protótipo do *PairExtractor*, foi implementado a leitura de arquivos no formato de planilha Excel. A forma de como será efetuada a leitura do arquivo, indicando quais colunas serão levadas em consideração para análise pode ser definida através de parâmetros de configuração do sistema. O sistema pode ser facilmente extendido para dar suporte a arquivos de formatos diferentes.

Um exemplo de planilha Excel contendo uma base de opiniões é ilustrado através da Figura 3.2

Produto	Tipo	Cons	PROS	Review	Stars
Ratatouille	movie	Cons: none	Pros: Great	oii madd dudes. this is a pretty radical movie you have made here. we give you a hi5.	5.0
Ratatouille	movie			There isn't really much that needs to be said about this film other than it is fab, funny and such a hoot. An	5.0
Ratatouille	movie			this movie was alsome i mean it is 95 % better than sharktale and those other movies.that movie was so c	4.0
Ratatouille	movie			While this was an enjoyable movie experience, it hardly warrants a 96. The plot became too mechanical, t	3.0
Ratatouille	movie			Well, I liked the previous Pixar movies pretty much. But I cried like a baby on this one. And I am a 28 yea	5.0
Ratatouille	movie			I bought this movie of course for my grand daughter, my first impression was that she would not really war	5.0

Figura 3.2 Arquivo com opiniões

Fonte: Elaboração Própria

3.5. Pré-Processamento

O primeiro módulo do sistema é responsável pelo pré-processameto das opiniões, de forma a produzir um arquivo de saída que será utilizado pelo módulo seguinte, o módulo de extração de pares. A seguir são descritas as principais tarefas deste módulo.

3.5.1. Segmentação das opiniões

O pré-processamento inicia-se através da leitura do arquivo contendo a base de opiniões. Cada linha do arquivo de opiniões contém uma opinião que pode ser composta por uma o mais sentenças.

Estas opiniões são então subdivididas em sentenças. Esta divisão será útil no processo de identificação de características não frequêntes, que será descrito com maior detalhes na Seção 3.6.

Para o processo de segmentação das opiniões em sentenças, são utilizados os caracteres de pontuação final (ponto, ponto e virgula, interrogação e exclamação).

De forma a evitar que abreviações sejam segmentados indevidamente, o sistema conta com uma base de abreviações, nas quais as abreviações contidas em uma opinião são substituídas pela sua forma extensa correspondente.

No protótipo desenvolvido, a lista de abreviações foi criada de forma manual, utilizando como base uma lista de abreviações descritas no dicionário da língua inglesa Oxford [9].

O Exemplo 3.2 ilustra a segmentação de uma frase simples, bem como a substituição de abreviações pela sua forma extensa.

Antes:

"Yesterday I watched Dr. Dolittle. It's a funny film."

Depois:

"Yesterday I watched Doctor Dolittle."
"It's a funny film."

Exemplo 3.2 Segmentação de uma opinião

3.5.2. POS-Tagging

Após a segmentação das opiniões, a etapa seguinte consiste em realizar o *POS-Tagging* de cada palavra, em cada sentença. Este procedimento consiste em identificar a classe gramatical de cada palavra de acordo com o sua definição e contexto [32].

Para realizar o *POS-Tagging*, foi selecionado o Tree-Tagger [27], uma ferramenta que realiza o *POS-Tagging* em diversos idiomas. Esta ferramenta recebe como entrada um arquivo com um conjunto de palavras, separadas por quebras de linha e fornece como saída o lema e a classe gramatical de cada palavra.

A Tabela 3.1 ilustra a saída produzida pelo Tree-Tagger após fornecer como entrada as sentenças do Exemplo 3.2.

Palavra	Classe gramatical	Lema
Yesterday	RB	Yesterday
I	PP	I
watched	VVD	watch
Doctor	NP	Doctor
Dolittle	NP	Dolittle
.	SENT	.
It	PP	it
is	VBZ	be
a	DT	A
funny	JJ	funny
film	NN	film
.	SENT	.

Tabela 3.1 POS-Tagging com o Tree-Tagger

Fonte: Elaboração Própria

Para permitir a identificação da segmentação do conjunto de opiniões em sentenças, é inserida uma uma palavra especial ".EOS." (*end of sentence*)

como indicador do fim de uma frase, de forma a ser utilizado no módulo seguinte de extração de pares.

De forma a aperfeiçoar o desempenho do Tree-Tagger, uma base de contrações e uma base de gírias podem ser utilizadas pelo sistema.

Contrações são utilizadas com grande frequência no inglês e, na presença de contrações, foi verificado experimentalmente que a classificação do Tree-Tagger não se comporta de forma adequada. O mesmo ocorre com gírias comumente utilizadas no vocabulário informal da internet.

As bases de contrações e de gírias funcionam de forma similar à base de abreviações, substituindo cada contração pela sua forma completa e cada gíria pelo vocábulo formal correspondente.

Ambas as bases foram criadas de forma manual no protótipo desenvolvido do *PairExtractor*. A lista de contrações foram obtidas através da gramática da língua inglesa Cambridge Grammar of English [2]. A lista de gírias foi criada com base em algumas gírias descritas no site noslang.com [19].

O Exemplo 3.3 ilustra uma possível entrada para o módulo de pré-processamento e a saída correspondente.

Entrada:

"*Fantastic Mr. Fox is gr8! Don't miss it!*"

Saída:

Palavra	Classe Gramatical	Lemma
Fantastic	JJ	fantastic
Mister	NP	Mister
Fox	NP	Fox
is	VBZ	be
great	JJ	great
!	SENT	!
Do	VVP	do
not	RB	not
miss	VV	miss
it	PP	It
!	SENT	!

Exemplo 3.3 Saída gerada pelo modulo de pré-processamento

Podemos observar através da análise do Exemplo 3.3, que a abreviação "Mr." foi traduzida para sua forma extensa "Mister", a gíria "gr8" foi traduzida para o vocábulo correspondente "great" e a contração "Don't" foi traduzida para sua forma completa "Do not". Após estes procedimentos, a frase foi etiquetada através da utilização do Tree-Tagger.

3.6. Extração de Pares

O módulo seguinte ao de pré-processamento consiste no módulo de extração de pares. Este é o módulo central do *PairExtractor*, onde as características são extraídas de fato.

Pesquisas indicam que quando usuários comentam sobre as características de um objeto, a maior parte das características descritas são substantivos e o vocabulário que eles utilizam geralmente converge [15].

As caracteristicas mais frequentes de um objeto são normalmente as mais importantes para um determinado domínio. Porém, existem algumas caracteristicas, com uma baixa frequência de avaliações, que podem ser interessantes para algumas aplicações.

Normalmente, para cada característica existem palavras opinativas próximas que as modificam de forma a expressar o ponto de vista do portador de opinião. Estas palavras opinativas geralmente são adjetivos ou advérbios. Desta forma, é importante a identificação de palavras opinativas de forma a encontrar características infrequentes.

Um dos pontos de originalidade deste trabalho, consiste na saída produzida pelo *PairExtractor*, que diferentemente dos outros trabalhos na literatura, não consiste apenas em uma lista de características, mas em uma lista de pares contendo características e palavras opinativas.

Como dito na Seção 3.1, estes pares representam uma melhoria em potencial para a etapa de classificação de sentimentos, pois a precisão alcançada será bem maior do que se fossem utilizados apenas características isoladas e palavras opinativas com polaridade determinadas a priori. O processo de extração de pares é descrito a seguir.

3.6.1. Extração dos Pares Frequentes

No método de extração de características descrito por Siqueira [26], a primeira etapa de extração consiste na identificação de um conjunto inicial de características candidatas, que contém os substantivos mais frequentes do *corpus* de entrada.

No método proposto neste trabalho, a etapa inicial da extração de pares ocorre de forma similar. Esta etapa consiste na identificação das características frequentes em um conjunto de opiniões e as possíveis palavras opinativas relacionadas a cada uma destas características.

O módulo de extração de pares utiliza como entrada o arquivo gerado pelo Tree-Tagger, descrito na seção anterior. A Tabela 3.2 contém algumas

das Tags geradas pelo TreeTagger. Estas serão as tags utilizadas pelo *PairExtractor* para a extração de pares [24].

Tag	Descrição
NN	Substantivo
NNS	Substantivo plural
NP	Nome próprio
NPS	Nome próprio plural
JJ	Adjetivo
JJR	Adjetivo comparativo
JJS	Adjetivo superlativo
RB	Advérbio
RBR	Advérbio comparativo
RBS	Advérbio superlativo

Tabela 3.2 Tags utilizadas para extração de pares

Fonte: Elaboração Própria

Inicialmente, todos os substantivos encontrados são armazenados em uma lista. Para cada substantivo, existe um contador e uma lista de palavras opinativas associadas, que podem ser adjetivos, advérbios, verbos ou outro substantivo.

Para cada ocorrência de um mesmo substantivo, são verificados se existem novas palavras opinativas associadas a ele, e caso existam, são adicionadas à lista de palavras opinativas. O contador também é incrementado para cada nova ocorrência.

Normalmente, as palavras opinativas são identificadas através de sua classe gramatical e proximidade com os substantivos candidatos a características. Neste trabalho, contudo, as palavras opinativas são identificadas com base em regras gramaticais. Este é um outro ponto de originalidade neste trabalho, pois na literatura relacionada não foi verificada nenhuma técnica que utilizasse uma abordagem similar.

As regras gramaticais utilizadas para a identificação de palavras opinativas foram elaboradas com base na gramática da língua inglesa Cambridge Grammar of English [2].

Devido à complexidade inerente ao entendimento da linguagem humana, não foi possível listar todas as construções possíveis onde possam existir palavras opinativas, mas foi verificado experimentalmente que as regras identificadas neste trabalho satisfazem a grande maioria das situações.

A Tabela 3.3 descreve as regras que foram aplicadas para a identificação das palavras opinativas com base nos substantivos identificados como candidatos a características. As classes sublinhadas representam as características e as classes em negrito representam as palavras opinativas.

Palavras Opinativas Anteriores à Característica
(1) Substantivo + Substantivo
(2) Verbo + Substantivo
(3) Adjetivo + Substantivo
Palavras Opinativas Posteriores à Característica
(4) Substantivo + **Adjetivo**
(5) Verbo + Substantivo + **Adjetivo**
(6) Substantivo + Verbo + **Adjetivo**
(7) Verbo + Substantivo + Advérbio + **Adjetivo**
(8) Substantivo + Verbo + Advérbio + **Adjetivo**
(9) Substantivo + Verbo + Advérbio + Advérbio + **Adjetivo**

Tabela 3.3 Regras gramaticais

Fonte: Elaboração Própria

Para o processo de casamento de padrão de forma a identificar as palavras opinativas foram desconsiderados os *stopwords*, que são palavras de classes gramaticais consideradas irrelevantes para o desenvolvimento deste trabalho, como artigos, pronomes, preposições, etc.

O *token* especial indicador de fim de sentença ".EOS.", foi utilizado nesta etapa de forma a indicar os limites à esquerda e à direita da característica, onde seriam buscadas as palavras opinativas, ou seja, dentro de uma sentença.

O Exemplo 3.4 ilustra a identificação de palavras opinativas através do uso das regras gramaticais.

.EOS.
The (DT)
phone (NN) is (VBZ) **great** (JJ) **Regra 6**
but (CC) the (DT)
camera (NN) is (VBZ) really (RB) **bad** (JJ) **Regra 8**
.(SENT)
.EOS.

Exemplo 3.4 Utilização de regras gramaticais para identificação de palavras opinativas

Podemos observar que no trecho "phone is great" foi utilizado a regra 6 para a identificação da palavra opinativa "great" e no trecho "camera is really bad" foi utilizado a regra 8 para a identificação da palavra opinativa bad.

Ao fim do processo de identificação dos substantivos candidatos à características, de suas respectivas contagens e palavras opinativas associadas, o último passo da etapa de extraçao dos pares frequentes consiste em selecionar os substantivos com maiores chances de serem características baseado em um *threshold* de relevância. Como este *threshold* pode variar de aplicação para aplicação, ele deve ser fornecido ao sistema através de um parâmetro de configuração.

Uma nova lista é então criada com base nos *t* substantivos de maior ocorrência, onde *t* é o *threshold* em valores percentuais.

3.6.2. Extração dos Pares Infrequentes

A etapa seguinte do processo de extração de pares consiste em identificar pares com características pouco frequentes, que vão então dar origem a pares infrequentes na lista de pares candidatos. Como dito anteriormente, para determinadas aplicações, a identificação de características infrequentes pode ser importante, devido à sua relevância para o domínio em questão.

A identificação dos pares infrequentes é realizada através das palavras opinativas associadas aos pares frequentes. Estudos indicam que as palavras opinativas são bons indicadores de presença de características verdadeiras [15].

No método de extração de características descrito por Siqueira [26], esta etapa consiste em selecionar os substantivos que ocorrem próximos a qualquer palavra opinativa, que ocorreu próxima a um outro substantivo participante da lista de características frequentes.

No PairExtractor, a identificação dos pares infrequentes é realizada através do acesso à lista de pares frequentes e para cada palavra opinativa distinta são buscados novos substantivos candidatos à características através da aplicação das regras gramaticas definidas anteriormente.

No final desta etapa, os pares frequentes e infrequentes são mesclados, constituindo em uma única lista de pares, com as características candidatas e suas respectivas palavras opinativas.

3.6.3. Mapeamento dos Indicadores de Características

Além das características explicitamente contidas no texto na forma de um substantivo, existem ainda as características que estão implicitamente em um texto opinativo. Estas características podem ser identificadas através de indicadores de características.

Os indicadores de características são palavras que, quando presentes em uma opinião, indicam a presença de uma característica implicitamente contida em uma sentença. Porém, este processo de mapeamento deve ser feito de com cuidado, pois existem indicadores de características que variam de acordo com o contexto.

O Exemplo 3.2 ilustra uma situação onde temos um indicador de característica que, dependendo do contexto, indica diferentes características implícitas.

*"The car is **heavy**."* (peso) *"The traffic is **heavy**."* (fluxo)

Exemplo 3.5 Indicador de característica dependente de contexto

A criação e o uso correto desses mapeamentos é um problema ainda em aberto na Análise de Sentimentos. Poucas pesquisas foram feitas sobre este tópico.

No método descrito por Siqueira [26], o mapeamento dos indicadores de características consiste em uma etapa distinta realizada após a etapa de identificação das características infrequentes, onde possíveis características implícitas são mapeadas através dos indicadores de características.

No *PairExtractor*, o mapeamento de características não é tido como uma etapa, mas como um processo de otimização, e é realizado durante as etapas de extração dos pares frequentes e infrequentes, de forma a conservar as palavras opinativas associadas às características implícitas.

O mapeamento dos indicadores em características é realizado através da utilização de uma base de mapeadores de características criada de forma manual.

3.7. Filtro de Pares

Os módulos anteriores descreveram o processo de identificação de pares (característica, palavra opinativa) com características frequentes, infrequentes ou implícitas. Porém, algumas das características previamente identificadas podem não ser características verdadeiras para um determinado domínio.

Uma das formas de verificar a relevância de uma palavra a um domínio, consiste na utilização de técnicas probabilísticas para determinar a probabilidade de duas palavras estarem juntas e a probabilidade destas estarem separadas.

O Método descrito por Siqueira [26] utiliza uma medida denominada de PMI-IR (*Pointwise Mutual Information*) [34] para realizar o processo de filtro de

características. Esta técnica foi comparada com uma outra denominada de NGD (*Normalized Google Distance*) [3], onde a NGD obteve melhores resultados. Este estudo será descrito com maior detalhes no Capítulo 4.

A técnica selecionada para o filtro dos pares não relacionados ao domínio, consiste no trabalho descrito por Cilibrasi e Vitányi [3]. Esta técnica é baseada na complexidade de Kolmogorov [18] e na distância normalizada da informação [29].

A complexidade de Kolmogorov, consiste na medida do tamanho do menor programa capaz de computar uma quantidade de informação passada como entrada. A distância normalizada da informação utiliza a complexidade de Kolmogorov para definir uma medida universal de distância entre objetos.

Ambas as métricas, são tidas como não computáveis na teoria computacional atual. Portanto, são utilizadas aproximações destas métricas no trabalho descrito por Cilibrasi e Vitányi.

Com base nas teorias préviamente descritas e utilizando o Google como motor de busca, os autores definem um método formal de cálculo de relevância de palavras ao qual eles denominam de distância Google normalizada (*Normalized Google Distance*), que é calculado através da seguinte fórmula:

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

Onde f(x) é a quantidade de resultados retornados para a pesquisa de uma palavra “x”, f(y) é a quantidade de resultados retornados para a pesquisa de uma palavra “y”, f(x,y) é a quantidade de resultados retornados para a pesquisa de ambas as palavras e N é um fator normalizador.

Os autores utilizam como fator normalizador, o número total de páginas indexadas pelo motor de busca, mas especificam que pode ser utilizado qualquer valor suficientemente maior que qualquer f(x) como fator normalizador.

A vantagem de se utilizar o total de páginas indexadas pelo motor de busca é que pode-se obter resultados equivalentes para a relevância de duas palavras ao longo do tempo. O número de páginas indexadas por cada motor de busca podem ser estimados através dos dados contidos no site WorldWideWebSize.com [35].

Apesar do método descrito ter utilizado o motor de busca do Google, pode-se utilizar outros motores de busca para o cálculo de relevância de palavras. No protótipo desenvolvido do *PairExtractor* pode-se utilizar o Google ou o Bing como motores de busca no cálculo de relevância.

O processo de filtro de pares consiste em calcular, para cada par, a relevância da característica candidata definida no par com relação ao domínio da aplicação. Como o nível de relevância aceito pode variar dependendo da aplicação, o *threshold* que delimita a partir de qual nível de relevância as palavras serão aceitas, é passado como parâmetro de configuração para o sistema. O fator normalizador também é passado como parâmetro.

3.8. Considerações Finais

Foi apresentado neste capítulo, o método proposto para extração de pares, o *PairExtractor*. Cada módulo que compõe o sistema foi descrito em detalhes, bem como as técnicas e procedimentos adotados, apresentando quando necessárias, as justificativas para a utilização das mesmas e ressaltando os pontos de originalidade deste trabalho.

No capítulo seguinte será mostrado os experimentos e resultados obtidos através de testes realizados em algumas técnicas empregadas no *PairExtractor* e no sistema como um todo.

4. Experimentos e Resultados

Neste capítulo serão mostrados a metodologia proposta para a realização de testes no *PairExtractor*, bem como os resultados alcançados e a sua avaliação.

A Seção 4.1 descreve a metodologia adotada para a realização de testes no *PairExtractor*, definindo o *corpus*, medidas utilizadas para a avaliação e as configurações de entrada. A Seção 4.2 contém os estudos realizados para a avaliação das técnicas de cálculo de relevância. A Seção 4.3 contém a análise das características extraídas pelo *PairExtractor*. A Seção 4.4 contém a análise da técnica baseada em regras gramaticais para identificação de pares. Por fim, a Seção 4.5 apresenta as considerações finais deste capítulo.

4.1. Metodologia

Para a realização dos experimentos, foi necessário a extração de um conjunto de opiniões para posteriormente criar um *corpus* de testes. Para este propósito, foram extraídas cerca de 30000 opiniões de diferentes domínios através de um Web *crawler*, que coletou um conjunto de *reviews* disponíveis no site do Google Products [5].

Após a extração das opiniões, foram escolhidas aleatoreamente 200 opiniões nos domínios de filmes e aparelhos celulares para a criação de 2 *corpus* de testes, contendo 100 opiniões cada.

Para cada opinião contida nos *corpus* de testes, foram manualmente identificadas as características (explícitas e implícitas) presentes, bem como as palavras opinativas relacionadas a cada uma destas características. No domínio de aparelhos celulares foram identificadas 71 características e 159 pares possíveis. No domínio de filmes foram identificadas 64 características e 136 pares.

4.1.1. Métricas

Para a avaliação dos testes realizados no *PairExtractor*, foram utilizadas 3 métricas: Precisão, Cobertura e F-measure. Estas medidas são utilizadas com frequência para avaliar técnicas da Análise de Sentimentos [26].

A precisão consiste na fração de instâncias relevantes em um determinado conjunto [33]. A precisão pode ser definida a partir da seguinte fórmula:

$$\text{Precisão} = \frac{tp}{tp + fp}$$

Onde tp (*true positive*) consiste nas instâncias relevantes e fp (*false positive*) consiste nas instâncias irrelevantes.

A cobertura consiste na proporção de instâncias relevantes contidas em um determinado conjunto, com relação ao número total de instâncias relevantes possíveis [33]. É definida através da seguinte fórmula:

$$\text{Cobertura} = \frac{tp}{tp + fn}$$

Onde tp (*true positive*) consiste nas instâncias relevantes contidas no conjunto e fn (*false negative*) consiste nas instâncias relevantes que não foram inseridas no conjunto.

A terceira medida, denominada de F-measure consiste na média harmônica da precisão e cobertura [33], definida através da seguinte fórmula:

$$\text{F-measure} = \left(\frac{2 \times \text{Precisão} \times \text{Cobertura}}{\text{Precisão} + \text{Cobertura}} \right)$$

4.1.2. Configurações

Com o objetivo de melhor analisar o *PairExtractor*, foram utilizadas 4 configurações diferentes de cada *corpus* de testes. A Tabela 4.1. ilustra as configurações utilizadas para os testes:

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 4.1 Configurações de testes

Fonte: Elaboração Própria

Em cada *corpus* de testes foram utilizados dois valores de *threshold* de frequência e dois valores de *threshold* de relevância. Como vistos no capítulo anterior, o *threshold* de frequência é utilizado para definir a quantidade de pares frequentes extraídos e o *threshold* de relevância é utilizado para determinar com que grau de rigidez os pares são filtrados com base no domínio.

4.2. Análise de Técnicas para Cálculo de Relevância

Nesta seção será mostrado em maior detalhes, o experimento realizado para determinar a técnica de cálculo de relevância mais eficiente de forma a ser utilizada pelo *PairExtractor* no filtro de pares com base no domínio.

As técnicas de relevância avaliadas neste experimento foram a NGD (*Normalized Google Distance*) [3], descrita em detalhes capítulo anterior, e a PMI-IR (*Pointwise Mutual Information*) [34], utilizada pelo *WhatMatter* proposito por Siqueira [26].

PMI-IR consiste em uma medida que indica o quão relacionadas estão duas palavras, através da utilização da probabilidade de encontrar juntas estas palavras e a probabilidade de encontrá-las de forma separada. O PMI-IR é calculado através da seguinte fórmula:

$$PMI_{IR}(p1, p2) = log_2 \left(\frac{Hits(p1 \wedge p2)}{Hits(p1) * Hits(p2)} \right)$$

Foram utilizadas as métricas definidas na seção anterior de forma a avaliar o desempenho de ambas as técnicas.

O experimento foi conduzido da seguinte forma: Foram extraídos pares frequêntes e infrequêntes utilizando como entrada o corpus de teste de aparelhos celulares. Após isto, cada técnica foi utilizada para filtrar os pares com base na sua relevância com o domínio. O *threshold* de cada técnica foi definido experimentalmente.

As medidas foram cálculadas utilizando como base o resultado esperado, definido de forma manual. A Tabela 4.2 ilustra os resultados obtidos com cada técnica:

Técnica	Precisão	Cobertura	F-measure
NGD	0.722	0.433	0.541
PMI-IR	0.272	0.200	0.230

Tabela 4.2 Resultado dos testes de técnicas de relevância

Fonte: Elaboração Própria

Através da análise da Tabela 4.2, podemos observar que a utilização do NGD foi claramente mais eficiente em todos os aspectos. Além de filtrar uma quantidade maior de pares nao relacionados ao domínio, também descartou uma quantidade menor de pares relacionados ao domínio.

4.3. Análise da Extração de Características

O experimento com objetivo de analisar a qualidade das características extraídas pelo *PairExtractor* foi realizada através da utilização dos dois corpus de testes como entrada, cada um com 4 configurações possíveis.

O processo de cálculo das medidas foi realizado em três etapas do *PairExtractor*. A etapa 1 consiste nos resultados obtidos com a identificação dos pares frequêntes. A etapa 2, consiste nos resultados obtidos após a etapa 1 e juntamente com a identificação dos pares infrequentes. A etapa 3 consiste nos resultados obtidos após as etapas 1 e 2, com o uso do filtro de pares.

A Tabela 4.3 ilustra os resultados obtidos com cada configuração e após cada etapa:

	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 4.3 Resultado dos testes de extração de características

Fonte: Elaboração Própria

Com base na análise da Tabela 4.3, podemos observar que a etapa 1 contém os índices mais altos de precisão e os mais baixos de cobertura. Isto é um fato previsto, pois através da extração dos pares frequentes, o esperado é que se encontre uma proporção pequena de características falsas e também poucas características verdadeiras são extraídas.

Analisando os resultados da etapa 2, podemos observar um decréscimo da precisão e um aumento da cobertura e do F-measure. Isto deve-se ao fato de serem acrescentados novas características, tanto verdadeiras quanto falsas, através do processo de identificação dos pares infrequentes.

Pode-se verificar através dos resultados da etapa 3, um leve decréscimo na cobertura e um acréscimo na precisão e na F-measure, devido à utilização do filtro de pares com base na relevância com o domínio. Desta forma, são descartados boa parte dos pares que contém características falsas e eventualmente alguns pares que contém características verdadeiras.

Por fim, podemos observar que os resultados do domínio de aparelhos celulares foram melhores que os resultados do domínio de filmes. Isto pode ser

justificando, dentre outros fatores, devido ao fato de que o domínio de filmes possui um conjunto de características de menor convergência, variando sensívelmente de acordo com cada filme.

4.4. Análise da Extração de Palavras Opinativas

Após a análise da extração de características, uma outra análise importante a ser efetuada consiste na verificação da qualidade de identificação das palavras opinativas que estão relacionadas a uma determinada característica, definindo desta forma os pares a serem extraídos pelo *PairExtractor*.

O experimento realizado com o objetivo de avaliar a extração das palavras opinativas consistiu em selecionar 10 características de cada domínio e anotar manualmente o conjunto de palavras opinativas esperado para cada característica e comparar com o resultado obtido pelo *PairExtractor*.

A Tabela 4.3 ilustra os resultados obtidos para a identificação de palavras opinativas com base em regras gramaticais:

Domínio	Precisão	Cobertura	F-measure
Aparelhos Celulares	0.764	0.818	0.790
Filmes	0.708	0.809	0.755

Tabela 4.4 Resultado dos testes de extração de palavras opinativas

Fonte: Elaboração Própria

Pode-se observar que os resultados foram similares para ambos os domínios, isto deve-se ao fato de que, apesar da diferência de vocabulário utilizado, a estrutura gramatical utilizada pelos portadores de opinião converge.

Foi observado que parte dos erros associados ao método de extração de palavras opinativas deve-se a erros de classificação gramatical do Tree-Tagger. Pois, com algumas palavras classificadas de forma incorretas, a aplicação de regras gramaticais tende a encontrar algumas palavras opinativas falsas e também deixar de encontrar algumas palavras opinativas verdadeiras.

Portanto, caso seja utilizado um *POS-Tagger* com maior precisão, os resultados do método de identificação de palavras opinativas podem ser otimizados.

4.5. Considerações Finais

Neste capítulo foi apresentado todo o processo de experimentos realizados nos principais aspectos do *PairExtractor*, bem como os resultados obtidos. De forma geral, os resultados obtidos podem ser classificados como satisfatórios.

O capítulo seguinte apresentará a conclusão deste projeto indicando as principais contribuições deste trabalho para a Análise de Sentimentos, bem como as dificuldades encontradas e possíveis trabalhos futuros.

5. Conclusão

Este trabalho teve como objetivo a apresentação do *PairExtractor* para a extração de pares formados por características e palavras opinativas, baseado na solução para extração de características descrita por Siqueira [26].

Na Seção 5.1 serão mostradas as principais contribuições deste trabalho para a Análise de Sentimentos. A Seção 5.2 descreve as dificuldades encontradas durante o processo de concepção e desenvolvimento do *PairExtractor*. A Seção 5.3 descreve alguns possíveis trabalhos futuros que podem ser realizados com relação ao *PairExtractor*.

5.1. Principais Contribuições

Este trabalho realizou um levantamento teórico da Análise de Sentimentos, através da definição dos principais conceitos da área e da revisão literária de diversas técnicas propostas para cada etapa da Análise de Sentimentos.

A maior contribuição deste trabalho, consiste em propor um método automatizado e livre de domínio para a extração de pares compostos por características e palavras opinativas através da combinação de diversas técnicas e procedimentos.

Outra contribuição que pode ser citada, consiste na proposta de um método para a identificação de palavras opinativas através de regras gramaticais.

5.2. Dificuldades Encontradas

As principais dificuldades encontradas neste trabalho também são inerentes ao domínio de Processamento de Linguagem Natural, como processar textos com uma estrutura complexa ou com a utilização de um vocabulário informal.

Pode-se ainda citar o tempo disponível para a realização do trabalho como uma das dificuldades. Devido à restrição de tempo, o escopo do protótipo desenvolvido limitou-se à etapa de extração de características. Porém, apesar do prazo restrito, todos os objetivos traçados para este trabalho foram alcançados.

5.3. Trabalhos Futuros

Alguns pontos de melhoria foram identificados de forma a aperfeiçõar o presente trabalho, que não foram realizados neste trabalho devido, principalmente, às restrições de tempo e podem ser realizados em trabalhos futuros. As melhorias que podem ser efetuadas são as seguintes:

1. **Agrupamento de pares sinônimos**: Consiste em tratar as diversas características presentes nos pares que são sinônimos como uma única característica, de forma a compartilharem a contagem de ocorrências e as palavras opinativas associadas.

2. **Resolução de co-referência**: Consiste em tratar problemas de co-referência de forma a identificar qual característica é mencionada por um determinado pronome.

3. **Mapeamento de indicadores de características de forma automática:** Consiste em propor uma técnica independente de domínio com o propósito de identificar de forma automática indicadores de características e mapear para as características implícitas automaticamente.

4. **Correção ortográfica:** Consiste em realizar a correção ortográfica das opiniões de forma automática.

5. **Mapeamento das Abreviações:** Consiste em mapear as abreviações para sua forma extensa de forma automática.

6. **Combinação de POS-Taggers**: Consiste em utilizar outras ferramentas de *POS-Tagging* combinadas com o objetivo de melhorar a precisão da classificação gramatical das palavras.

Referências Bibliográficas

[1] Bruce, R.; and Wiebe, J.; Recognizing Subjectivity: A Case Study of Manual Tagging. Natural Language Engineering. 2000.

[2] Carter, R.; and McCarthy, M.; Cambridge Grammar of English. A Comprehensive Guide. Spoken and Written English Grammar and Usage. Cambridge University Press. 2006.

[3] Cilibrasi, R.; and Vitányi, P.; The Google Similarity Distance. In IEEE ITSOC Information Theory Workshop. 2005.

[4] Ding, X.; Liu, B.; and Yu, P.; A holistic lexicon-based approach to opinion mining. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM). 2008.

[5] Google; Google Products. Disponível em http://www.google.com/prdhp. Último acesso em 10 de Dezembro de 2011.

[6] Goldberg, A.; Zhu, X.; and Wright, S.; Dissimilarity in graph-based semisupervised classification. In Artificial Intelligence and Statistics (AISTATS). 2007.

[7] Hu, M.; and Liu, B.; Mining and summarizing customer reviews. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177, New York, NY, USA. 2004. ACM.

[8] Hu, M.; and Liu, B.; Mining opinion features in customer reviews. In AAAI'04: Proceedings of the 19th national conference on Artifical intelligence, pages 755–760. AAAI Press. 2004.

[9] Indiana University; The Oxford English Dictionary: List of Abbreviations. 1996. Disponível em http://www.indiana.edu/~letrs/help-services/QuickGuides/oed-abbr.html. Último acesso em 06 de Dezembro de 2011.

[10] Junior, A.; Detecção Automática de Subjetividade: Aprendizagem de Máquina Versus Ferramentas Linguísticas. Dissertação (Graduação em Ciência da Computaçao) - Centro de Informática/UFPE. Recife. 2011.

[11] Koblitz, L.; Ambiente de Análise de Sentimentos Baseado em Domínio. Dissertação (Doutorado em Engenharia Civil) - UFRJ. Rio de Janeiro. 2010.

[12] Kuo, F.; An investigation of effort-accuracy trade-off and the impact of self-efficacy on Web searching behaviors. Decision Support Systems, v.37 n.3, pp.331-342. 2004.

[13] Hatzivassiloglou, V.; and Wiebe, J.; Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING). 2000.

[14] Liu, B.; Hu, M.; and Cheng, J.; Opinion observer: Analyzing and comparing opinions on the web. In WWW '05: Proceedings of the 14th international conference on World Wide Web, pages 342–351, New York, NY, USA. 2005. ACM.

[15] Liu, B.; Sentiment Analysis and Subjectivity. In Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F.J. Damerau, Editors. 2010.

[16] Liu, B.; Sentiment Analysis and Opinion Mining Tutorial. Given at AAAI-2011. San Francisco, USA. 2011.

[17] Liu, B.; Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, The First Edition (2006) and Second Edition (2011). 2006 and 2011: Springer.

[18] Li, M.; and Vitányi, P.; An Introduction to Kolmogorov Complexity and Its Applications, 2nd Ed., Springer-Verlag, New York. 1997.

[19] NoSlang; Internet Slang Dictionary & Translator. Disponível em: http://www.noslang.com/dictionary/. Último acesso em 05 de Dezembro de 2011.

[20] Pang, B.; and Lee, L.; Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, v.2 n.1-2, p.1-135. 2008.

[21] Popescu, A.; and Etzioni, O.; Extracting product features and opinions from reviews. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2005). 2005.

[22] Qiu, G.; Liu, B.; Bu, J.; and Chen, C.; Expanding domain sentiment lexicon through double propagation. In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2009). 2009.

[23] Read, J.; Recognising Affect in Text using Pointwise-Mutual Information. 2004.

[24] Santorini, B.; Part-of-Speech Tagging Guidelines for the Penn Treebank Project. 1991.

[25] Silva, N.; BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião. Dissertação (Graduação em Ciência da Computação) – Centro de Informática/UFPE. Recife. 2010.

[26] Siqueira, H.; WhatMatter: Extração e visualização de características em opiniões sobre serviços. Dissertação (Mestrado em Ciência da Computaçao) - Centro de Informática/UFPE. Recife. 2010.

[27] Schmid, H.; Treetagger - a language independent part-of-speech tagger. Disponível em: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger. Último acesso em 08 de Dezembro de 2011.

[28] Universität Karlshure; Sentiment Classification bibliography. Disponível em: http://liinwww.ira.uka.de/bibliography/Misc/Sentiment.htm . Último acesso em 18 de Novembro de 2011.

[29] Vitányi, P.; Balbach, J.; Cilibrasi, R.; and Li, M.; Nomalized Information Distance. In Information Theory and Statistical Learning, Eds. M. Dehmer, F. Emmert-Streib, Springer-Verlag, New-York. 2008.

[30] Wiebe, J.; Opinion Mining bibliography. Disponível em: http://www.cs.pitt.edu/~wiebe/subjectivityBib.html. Último acesso em 15 de Novembro de 2011.

[31] Wiebe, J.; Bruce, R.; and O'Hara, T.; Development and Use of a Gold Standard Data Set for Subjectivity Classifications. In Proc. of ACL'99. 1999.

[32] Wikipedia; Part-of-speech tagging. Disponível em: http://en.wikipedia.org/wiki/Part-of-speech_tagging. Último acesso em 08 de Dezembro de 2011.

[33] Wikipedia; Precision and recall. Disponível em: http://en.wikipedia.org/wiki/Precision_and_recall. Último acesso em 11 de Dezembro de 2011.

[34] Wikipedia; Pointwise mutual information. Disponível em: http://en.wikipedia.org/wiki/Pointwise_mutual_information. Último acesso em 12 de Dezembro de 2011.

[35] WorldWideWebSize; The size of the World Wide Web (The Internet). Disponível em: http://www.worldwidewebsize.com. Último acesso em 12 de Dezembro de 2011.

[36] Yu, H.; and Hatzivassiloglou, V.; Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2003.

[37] Zhang, L.; and Liu, B.; Extracting and Ranking Product Features in Opinion Documents. Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010), August 23-27, Beijing, China.

Apêndice

A – Amostras da Base de Abreviações

1. Acct.	Account
2. Corp.	Corporation
3. Dept.	Department
4. Dr.	Doctor
5. Educ.	Education
6. Eng.	English
7. FAQ.	Frequently asked question
8. Ind.	Industry
9. Irreg.	Irregular
10. Mem.	Memory

B – Amostras da Base de Contrações

1. Aren't	Are not
2. Can't	Can not
3. Doesn't	Does not
4. Hadn't	Had not
5. I'll	I will
6. Mustn't	Must not
7. Needn't	Need not
8. She's	She is
9. That's	That is
10. Wasn't	Was not

C – Amostras da Base de Gírias

1. ASAP	As soon as possible
2. BST	best
3.BTFL	beautiful
4. BTW	By the way
5. DUNNO	Do not know
6. GR8	Great
7. PPL	people
8. TBH	to be honest
9. WANNA	want to
10. WUD	would

D – Amostras dos Indicadores de Características

1. Beautiful	Appearance
2. Cheap	Price
3. Expensive	Price
4. Large	Size
5. Light	Weight
6. Long	Duration
7. Money	Price
8. On Time	Delivery
9. Pocket	Size
10. Ugly	Appearance

E – Amostras do Corpus de Aparelhos Celulares

1. Great phone for the money. Easy to program and use. Clear sound, good enough camera for on the go pictures. Battery last about 3 days on standby. I'll recommend this phone ...

2. I have had the phone for slightly more than a month and had to go out and get a new battery yesterday. I guess this is a reminder that there is no such thing as a deal Let the buyer beware.

3. I love bluetooth, but every so often I find my 2.4 Ghz wireless interferes with my Razr bluetooth connection. This is the answer. I just wish there were some kind of ear loop to go with it.

4. I purchased this battery 4 days ago. My old battery had never used over 1 day. It made me to carry charger every time. Now, this new battery can be used for 2 days to 3 days. It is really convenient for me.

5. The case fell apart (cracked) about three days after I put it on. I only have the bottom part on it now because the other piece fell apart and it would no longer stay on.

6. I only have one issue with the phone and it is the battery life. The battery goes dead way too fast even when you don't use it much.

7. Although it is cute, this case is too small and is nearly impossible to remove once put on the phone. The plaid pattern also chips terribly. Do not buy this case.

8. The phone works great, but the battery will not fully charge.

9. This is absolutely the best smartphone I have ever owned! I wish the camera had more options and the battery life was longer, but over all this phone is the best smart phone I have ever used! Along with the warranty that Best Buy offers at an affordable rate, this phone is a great deal.

10. Excellent phone! Sleek design. Combination of keyboard and touchscreen makes for an enjoyable experience. Easy navigation and quick response time. Trackpad is a great addition. Battery life is good. A little heavier than most of other phones I've owned, but worth it & not noticeable after awhile. Would 100% recommend this phone.

F – Amostras do Corpus de Filmes

1. This is one tremendous movie for young girls. It has a great story line, great actors and actresses, a beautiful fairy tale ending and encourages girls to be more active in their lives. It makes a strong case for a classic movie, mainly from the standpoint of integrating fairy tale lore with today's modern era. Just a well written and produced movie that is great for all ages. Watch it again and again!

2. The movie was fun to watch for adults, teens, preteens,and toddler. It had a little something for everyone. Humor, fantasy, love, song, and heroics.

3. The first Harry Potter movie was awesome, but this one was a total dissapointment, much like the book was. The worst Harry Potter movie yet. The storyline bored me.

4. Great effects especially Dobby and Fawkes! I love the Dueling Club scene!! It is indeed a memorable one! Keep it up! Job well done for Daniel and Emma!!

5. You really must have this film regardless of age you will love the story, the really cute rats and and the wicked attention to detail, a simply wonderful piece…

6. Superb animation, entertaining characters and a fresh plot.. a fantastic, feel-good movie that brings joy to even the coldest hearts!

7. The premise is good and it taken to fruition quite well. The characters are all distinct and memorable and the plot makes sense.

8. I must admit that this film did have me on the edge of my seat because it was scary, fun and well written, however Shyamalan end this movie with the worst twist passable.

9. Great suspense movie by a good director. I liked this one better than The Sixth Sense. Acting is good and the theme is different even though it looks different.

10. Incredibly great movie, completely awesome, you gotta watch this one. A little more comedy than needed but still great.