BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião

Por
Nelson Gutemberg Rocha da Silva

Trabalho de Graduação

Recife, Dezembro 2010

Nelson Gutemberg Rocha da Silva

BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião

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

Orientadora: *Profa. Dra. Flávia Barros Almeida*

Recife, Dezembro 2010

A Neide, Severino e João Paulo

Agradecimentos

Agradeço primeiramente a Deus, aquele a quem devo tudo que tenho e sou. Ele que me guiou durante toda minha vida, e na qual tenho o maior orgulho em servi-lo.

Agradeço a minha família, que sempre me incentivou e depositou em mim muita confiança, em especial ao meu Pai, que antes de partir para o lado de Deus, deixou a maior herança que um filho poderia desejar: o seu amor e carinho; à minha mãe, Neide, e ao meu irmão João, que sempre me apoiaram em minhas decisões e sempre estiveram ao meu lado.

Agradeço também a todos os amigos que formei na universidade, amigos com quem compartilhei momentos difíceis e de alegria durante a graduação, em especial, Paulo Ricardo, Eduardo Gade, Felipe Kühner e João Rufino.

E por fim agradeço aos professores do Centro de Informática – UFPE, que contribuíram com minha formação acadêmica. Em especial à professora Flávia Barros, que me orientou nesse trabalho de graduação e me guiou para que possa obter êxito na conclusão de minha graduação.

A todos, o meu muito Obrigado.

Comece fazendo o que é necessário, depois o que é possível, e de repente você estará fazendo o impossível.

São Francisco de Assis

Resumo

A internet tem se tornado, a passos largos, a maior base de informação do mundo, principalmente quando surgiram as ferramentas de expressão de opiniões, como blogs, fóruns e redes sociais, que trazem consigo a experiência de milhões de usuário acerca dos mais diversos produtos.

Esse tipo de informação traz dados importantes para empresas que pretendem melhorar e divulgar seus produtos: as opiniões de seus clientes em relação a seus produtos e aos produtos da concorrência. De outro ponto de vista, as opiniões e experiências de outros usuários sobre alguma marca, produto ou serviço tornam-se de extrema importância na hora de tomar uma decisão de compra. Porém, toda essa informação está dispersa de forma não organizada por toda Web, e devido ao grande volume de informação, a busca pelas opiniões dos usuários torna-se uma tarefa impraticável. Estima-se que cerca de 75.000 blogs e 1,2 milhão de novas postagens são criadas por dia.

Nesse contexto, estão surgindo sistemas para tratar opiniões automaticamente, utilizando-se do conceito da área conhecida como *Análise de Sentimentos* (AS). A AS, entre outras tarefas, preocupa-se em classificar opiniões expressas em textos, com respeito a um determinado produto ou serviço, como positivas ou negativas.

O objetivo central deste Trabalho de Graduação foi minerar opiniões de usuários a partir das informações disponíveis na Web, utilizando para isso técnicas de Análise de Sentimentos.

Sumário

1. Introdução 1
1.1. Contexto 2
1.2. Objetivos e Solução Proposta 3
1.3. Organização do Trabalho 5
2. Análise de Sentimentos 6
2.1. Conceitos Básicos 6
2.1.1. Pointwise Mutual Information 9
2.2. Algumas Aplicações 10
2.3. Etapas da Análise de Sentimentos 10
2.3.1. Subjetividade 11
2.3.2. Extração de Características 11
2.3.3. Classificação 12
2.3.4. Visualização e Sumarização 13
2.4. Desafios e Limitações da Área 14
2.4.1. Identificação da Subjetividade 14
2.4.1. Classificação da Polaridade dos Adjetivos e Advérbios 15
2.4.1. Outros Desafios com Processamento de Texto 15
2.5. Conclusão 15
3. Classificação de Sentimentos 16
3.1. Motivação para trabalhar com classificação 16
3.1.1. Desafios e Limitações 16
3.2. O Sistema HowGood 17
3.2.1. Visão geral 17
3.2.2. Crítica 20
3.3. SentiWordNet 21
3.4. Considerações finais 22
4. BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião 23
4.1. Caracterização do Problema 23
4.2. Arquitetura Geral 24
4.3. Bases Externas 26
4.4. Módulo de Carga das Bases 26
4.5. Processamento 27
4.5.1. POS-Tagger 27
4.5.2. Termos 28
4.6. Polaridade 30
4.7. Classificação 31
4.8. Sumarização 34
4.8.1. Resultado Usuário 34
5. Experimentos e Resultados 35
5.1. Metodologia Para Experimentos 35
5.2. Resultados e Avaliação 36
6. Conclusão 40
6.1. Principais Contribuições 40
6.2. Dificuldades Encontradas 40
6.3. Trabalhos Futuros 41
REFERÊNCIAS BIBLIOGRÁFICAS 42

Lista de Figuras

Figura 2.1: ontologia de aspectos do objeto *Spettus* 8
Figura 2.2: Três celulares foram representados, cada um com sua cor. 14
Figura 3.1: Tabela de relevância de aspectos 18
Figura 3.2: Tabela de relevância de palavras opinativas como adjetivos 18
Figura 3.3: Tabela de relevância de palavras opinativas como advérbios 19
Figura 4.1: Arquitetura do BestChoice 24
Figura 4.2: Tela Principal. Consulta por Completo 25
Figura 4.3: Tela Principal. Consulta Parcial 25
Figura 4.4: Resultado da Carga 27
Figura 4.5: Resultado do Módulo Processamento 30
Figura 4.6: Tela para Alterar Polaridades 31
Figura 4.7: Tela de Seleção de Aspectos 31
Figura 4.8: Resultado mostrado para o usuário. 34

Lista de Tabelas

Tabela 1.1: Sequência extraídas 2
Tabela 1.2: Marcações 3
Tabela 4.1: Tabela com o resultado do processamento do Tree-Tagger 28
Tabela 4.2: Tabela com Advérbios e Locuções Adverbiais de Negação 32
Tabela 4.3: Cláusulas Adversativas 33
Tabela 5.1: Tabela parcial dos primeiros resultados de palavras opinativas 36
Tabela 5.2: Tabela com a as polaridades das palavras opinativas 37
Tabela 5.3: Resultado comparativo entre os aspectos do HowGood e BestChoice 38

1. Introdução

Milhares de anos foram precisos para o homem desenvolver a comunicação através da linguagem falada e escrita. Essa necessidade de comunicação fez com que o homem criasse também maneiras de se comunicar com as máquinas, através de linguagens sistemáticas e de comandos precisos, conhecidas como linguagens de máquina [BROOKSHEAR, 2005]. Porém, essa comunicação com máquina ainda é muito restrita quando a intenção é usar a linguagem que tanto tempo o homem demorou a desenvolver, a Linguagem (ou Língua) Natural.

O termo *Linguagem Natural* se refere à linguagem que os homens utilizam para se comunicar entre si. Atualmente, existe um ramo na computação que trata do Processamento da Linguagem Natural (PLN) [WITTMANN & RIBEIRO, 1998], do inglês *Natural Language Processing*, que é uma subárea de Inteligência Artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas naturais. Existem subáreas em PLN, que estudam tanto a conversão da linguagem de máquinas em Linguagens Natural, como a subáreas que estudam a conversão da linguagem Natural, em linguagens que podem ser interpretadas e processadas por programas de computador. O objetivo central da PLN é fornecer aos computadores a capacidade de compor e entender textos.

E com o notável crescimento da quantidade de informação disponível na Web nos últimos anos, o tratamento dessa informação de forma automática se torna indispensável para que se possa tirar o máximo de proveito dessa grande base de dados.

Além das notícias vinculadas nos portais de jornais, revistas e emissoras, outra forma de divulgar informação que tomou conta da Web, é através das ferramentas de expressão de opinião, como blogs, fóruns, sites de relacionamento, entre outros. Os internautas gastam em média 25% do tempo na Web nesse tipo de mídia [NIELSEN, 2009]. Os usuários disponibilizam nessas ferramentas uma importante base de informação, as opiniões e experiências acerca de um produto ou serviço. Essas opiniões podem ser de extrema importância para usuários e empresas, que pretendem comprar ou atualizar algum produto, ou contratar serviços de alguma empresa.

Apesar dessa informação estar disponível na Web, tratar essa informação de forma manual se torna uma tarefa impraticável. Seria preciso milhares de textos para poder avaliar a opinião geral acerca de um produto. Estima-se que cerca de 75.000 blogs e 1,2 milhões de novas postagens são criadas por dia [FRAZON & GONÇALVES, 2006]. Com esse aumento de informação postada por usuários, surgiu uma nova área de pesquisa relacionada a Processamento de Linguagem Natural, chamada de Mineração na Web e também conhecida como Análise de Sentimentos (AS) [LIU, 2010].

A Análise de Sentimento visa identificar as opiniões postadas por usuários na internet, avaliar, classificar como positivas, negativas ou neutras, e disponibilizar o resultado da análise de forma clara para o usuário final.

1.1. Contexto

AS é uma área que apenas recentemente vem crescendo no meio científico. Na década de 90 do século XX, foram publicados diversos artigos sobre AS, porém só nos últimos anos que houve uma grande quantidade de publicações [RODRIGUES, 2009].

O processo de AS é bastante complexo, e um sistema completo pode-se dividir em pelo menos quatro etapas [LIU et al., 2005]: Identificação de opiniões, Extrações de características, Classificação de sentimentos e Visualização e sumarização. Para cada uma dessas etapas, existem diversos trabalhos na literatura.

A maioria dos sistemas de AS identifica apenas a classificação de um texto, como positivo, negativo ou neutro, porém muita informação pode ser obtida com uma análise mais profunda. A classificação da opinião de um texto pode ser feita, basicamente, em três níveis: no nível de documento, no nível de sentença e no nível de aspecto [FERNANDES, 2010].

- No nível de documento, o objetivo é classificar a opinião global do usuário acerca do objeto sob análise. Neste caso, várias técnicas baseadas em aprendizado não supervisionado [LIU, 2010] [TURNEY, 2002] [SANTORINI, 1995] e supervisionado [KOTSIANTIS et al., 2007] [PANG & LEE, 2008] [LIU, 2006] [MAUÁ, 2009] podem ser usadas.
- No nível de sentença, a classificação não é feita a partir de todo o texto, mas sim de cada sentença pertencente ao texto, aumentando a granularidade da entrada, e, assim, melhorando o resultado final da classificação.
- No nível de aspecto, cada característica (i.e., aspecto) do produto a ser analisado é classificada, dando ao usuário final uma visão mais real e refinada das opiniões acerca do produto.

Turney em [TURNEY, 2002], realiza a classificação em nível de documento baseado em aprendizagem não supervisionada. Ele usa templates fixos de sintagmas para classificar os sentimentos, utilizando-se dos adjetivos e advérbios como fortes indicadores de sentimentos.

Inicialmente ele aplica o Pos-Tagger [SANTORINI, 1991] aplicando ao texto as marcações indicadas na tabela 1.2, depois localiza uma das sequências mostradas na tabela 1.1.

Primeira Palavra	**Segunda Palavra**	**Terceira Palavra**
JJ	NN ou NNS	Qualquer uma
RB, RBR ou RBS	JJ	Não NN, nem NNS
JJ	JJ	Não NN, nem NNS
NN ou NNS	JJ	Não NN, nem NNS
RB, RBR ou RBS	VB, VBD, VBN ou VBG	Qualquer uma

Tabela 1.1: Sequência extraídas

Marcação	Descrição
JJ	Adjetivo
NN	Substantivo
RB	Advérbio
VB	Verbo no Infinitivo
NNS	Substantivo no Plural
RBR	Advérbio Comparativo
RBS	Advérbio Superlativo
VBD	Verbo no Passado
VBN	Verbo No Particípio Passado
VBG	Verbo no Gerúndio

Tabela 1.2: Marcações

Após a extração dos termos encontrados na primeira e na segunda coluna da tabela 1.1, é aplicado o calculo do PMI, como será mostrado na seção 2.1.1, para os termos retirados da tabela 1.1. O calculo da orientação de cada um dos termos é dado a partir associação de cada uma delas com os termos "excellent", que possui uma conotação positiva, e o termo "poor", que possui uma conotação negativa. A orientação é calculada da seguinte forma:

$$oo(phrase) = PMI(phrase, excellent) - PMI(phrase, poor)$$

Após encontrar a orientação dos termos, é realizada uma média entre as orientações encontradas no texto, para se retirar a orientação do documento.

Já em [LIU, 2006], a classificação em nível de sentença é dividida em duas sub-tarefas:

1. Classificar a Subjetividade: detectar se a sentença trata de uma opinião ou de uma sentença informativa.
2. Classificar a Orientação: Verificar se a sentença exprime um sentimento positivo ou negativo em relação ao objetivo avaliado.

Em [FERNANDES, 2010] é proposto o HowGood, que realiza a classificação em nível de aspecto. Na seção 3.2 entraremos em mais detalhes em relação ao HowGood e a classificação em nível de aspecto.

Veremos a seguir os objetivos deste TG, bem como um resumo da solução proposta.

1.2. Objetivos e Solução Proposta

O objetivo principal deste Trabalho de Graduação foi construir um sistema de Análise de Sentimentos para Ferramentas de Expressão de Opinião no nível do aspecto.

Este TG teve por base os estudos apresentados em [FERNANDES, 2010], que oferece uma solução para classificação de opiniões no nível de aspecto. O autor desenvolveu o sistema protótipo *HowGood*, que implementa um classificador de opiniões para postagens do Twitter[1] no domínio de bares e restaurantes.

Como na maioria dos trabalhos que tratam da AS, algumas das fases de processamento do sistema *HowGood* são realizadas de forma semiautomática, ou seja, com a intervenção do usuário. Em particular, duas tarefas são realizadas de forma manual: a escolha dos aspectos (e.g., cerveja, estacionamento) e das palavras opinativas (e.g., quente, caro) a serem considerados; e a classificação prévia das palavras opinativas (e.g., caro (neg), quente (neg), queimada (pos), etc). Assim, o desempenho do sistema pode ser prejudicado caso o usuário não tenha tempo ou discernimento para realizar a escolha e a classificação dessas palavras.

Além disso, com o aumento da base de informação (os textos com as opiniões), novas palavras serão continuamente selecionadas para análise, o que pode tornar muito difícil o uso do sistema.

Este trabalho apresenta uma proposta de melhoria para o processo proposto em [Fernandes, 2000] através da automação de fases intermediária do processo de Análise de Sentimentos que requerem a intervenção do usuário no sistema HowGood. Em particular, o nosso foco foi na classificação automática das palavras opinativas.

Para este fim, o sistema protótipo *BestChoice* (proposto neste TG) utiliza-se de uma base de termos usada para Mineração de Opinião, conhecida como *SentiWordNet* [ESULI & SEBASTIANI, 2006]. O SentiWordNet atribui a cada termo contido na base WordNet[2] [FELLBAUM, 1998] três pontuações de sentimento: negatividade, positividade, e objetividade. O sentimento de objetividade está relacionado à neutralidade do termo na oração, ou seja, quanto maior a pontuação de objetividade menor é a caracteristica negativa e positiva da palavra.

Contudo, o SentiWordNet utiliza o idioma inglês na versão adotada neste TG. Assim sendo, foi necessário também criar um módulo que traduz os termos de português para inglês, a fim de então obter a classificação das palavras opinativas.

O uso do SentiWordNet, somado ao tradutor de termos, traz três contribuições principais ao processo de AS:

- Eliminam a necesisdade de o usuário selecionar previamente os termos que serão usados como palavras opinativas, uma vez que o SentiWordNet cobre um vocabulário vastíssimo de termos. Isso favorece a estensibildade do sistema.
- Eliminam a necesisdade de o usuário atribuir previamnete sentimento às palavras opinativas, simplificando o uso do sistema.
- Oferecem a possibilidade de se trabalhar com textos em vários idiomas, uma vez que a tradução automática de termos é uma tarefa relativamente simples.

Além dessas contribições mais gerais para o processo de AS, ressaltamos ainda uma contribuição mais particular, que se refere ao problema da classificação dos termos em suas classes gramaticais (do inglês, parts-of-speech - POS). De modo a melhorar a classificação realizada pelo POS-tagger *TreeTagger* usado em [FERNANDES, 2010], a seguinte técnica foi usada neste trabalho:

[1] http:// twitter.com
[2] http://wordnet.princeton.edu/

- Verificar a classe gramatical (lexical) do termo no WordNet, para corrigir um eventual erro de classificação pelo TreeTagger;

Essa melhora na classificação do Tree-Tagger foi necessária porque foi verificado que a classificação errada da classe gramatical de algmas palavras resultou em uma queda na precisão da classificação do SentiWordNet. Com a melhora na classificação da classe gramatical das palavras, o resultado se mostrou bem mais satisfatório, como será descrito no capítulo 6.

Nossos experimentos se inspiraram na mesma base de dados (*twitts*) utilizada em [FERNANDES, 2010], disponibilizada pelo autor, para possibilitar a comparação dos resultados de precisão da classificação de sentimento.

1.3. Organização do Trabalho

O trabalho em questão está divido em seis capítulos. Após o capítulo introdutório, mais cinco capítulos tratam da revisão bibliográfica e dos detalhes do sistema produzido e seus resultados:

- O capítulo 2 apresenta brevemente a área de Análise de Sentimentos. Serão aprensentados os principais conceitos, cada uma de suas etapas de processamento, e as técnicas mais conhecidas e utilizadas no mercado.

- No capítulo 3, é mostrado o estado da arte de uma das fases de AS, a Classifiação, que é o foco principal deste TG. Serão mostradas técnicas e pesquisas existentes, e os principais desafios da área de pesquisa. Veremos aqui, também o sistema HowGood [FERNANDES, 2010], e o SentiWordNet.

- No capítulo 4, é mostrado o processo de AS proposto neste trabalho, bem como o sistema *BestChoice*. Serão descritos os algorítimos e as técnicas usadas para melhorar o sistema HowGood.

- No capítulo 5, serão mostrados os resultados: a metodologia, experimentos e uma comparação dos resultados obitidos com os resultados obitidos em [FERNANDES, 2010].

- O capítulo 6 traz a conclusão do trabalho, fazendo uma análise dos resultados obitidos, relatando algumas possíveis melhorias futuras e os objetivos alcançados.

2. Análise de Sentimentos

Análise de Sentimentos (AS), também conhecido como Mineração de Opinião é uma área da computação bastante recente, que estuda a classificação das emoções e opiniões expressas em textos opinativos [LIU et al., 2009].

Essa é uma área de conhecimento recente que tem crescido muito nos últimos anos. Com o aumento de pessoas que têm acesso à internet e das ferramentas de expressão de opinião, os textos com as opiniões e experiência de usuários vêm se tornando a maior parte do conteúdo da Web, transformando as ferramentas de AS em um fator importante para se extrair informação desse emaranhado de textos. Segundo o *Gartner Group* (Empresa de Consultoria fundada em 1979 por Gideon Gartner), AS está entre as 10 tecnologias estratégicas para as corporações no próximo ano [IPNEWS, 2010]. A técnica de AS não só é bastante desafiadora, como também, muito útil na prática.

A AS já é realizada de forma manual, por usuários que pretendem obter informações, opiniões e experiência de outros usuários acerca de um produto. Muitos usuários já fazem pesquisas em blogs e fóruns, buscando essas opiniões antes de adquirir algum produto. Porém, essa prática não traz resultados satisfatórios, e muitas vezes não reais, por causa da pequena quantidade de informação na qual o usuário faz sua análise, principalmente pelo fato da opinião ser um conceito distorcido devido a sua carga de subjetividade [MAGALHÃES, 2009]. Ou seja, enquanto alguns usuários falam bem de um produto, outros usuários podem falar de forma negativa do mesmo produto. Logo, analisar várias opiniões por toda Web se torna importante para se ter uma visão geral da opinião dos usuários em torno do produto.

Outro grupo de usuários em potencial são as empresas que pretendem lançar um novo produto, ou uma nova versão de um já lançado no mercado. Para esse segundo grupo de usuários, a análise das opiniões dos consumidores é de extrema importância, pois um produto que atinja os gostos do consumidor pode ocasionar uma grande perda para empresa.

Neste capítulo, serão mostrados os conceitos básicos de AS, suas aplicações e alguns produtos que já estão no mercado, algumas limitações ainda não superadas nessa área, e as principais etapas para o problema de AS.

2.1. Conceitos Básicos

Análise de Sentimentos, por ser uma área bastante recente e ainda não se ter um padrão para os nomes usados na área, ganha diversos termos relacionados ao mesmo conceito. De modo a ficar mais claro o conteúdo em torno de AS, serão mostrados aqui alguns conceitos básicos da Análise de Sentimento.

A maioria dos conceitos usados nesse trabalho irá seguir o padrão proposto em [LIU, 2006], onde o autor se baseia nos conceitos mais comuns da literatura que trata de AS. O texto abaixo irá servir de base para exemplificar cada um dos conceitos mostrados:

> *"(1) Ontem comemoramos o aniversário do meu primo no Spettus. (2) (2.1) A comida do Spettus é maravilhosa, e (2.2) lá não é caro. (3) (3.1) A carne do Spettus é muito boa, mas (3.2) o sushi não é um dos melhores."*

Exemplo 2.1: comentário criado pelo autor.

Ao analisarmos o **exemplo 2.1**, o primeiro fator a analisarmos é a subjetividade do texto, ou seja, trata-se de um texto opinativo ou de um texto subjetivo. Sendo este um texto opinativo, as opiniões são extraídas do texto para serem analisadas.

Como a oração (1) se trata de um texto subjetivo, pois demonstra um fato ocorrido e não uma opinião acerca do objeto a ser analisado (Spettus), a oração seria descartada da AS. Já as orações (2) e (3), tratam claramente de uma opinião.

Outro fator importante ao analisarmos o texto, é o tipo de classificação a ser usada, pois dependendo da classificação usada os resultados podem ser diferentes para o usuário final. Em uma classificação em nível de texto, poderia se dizer que a opinião do texto é positiva, considerando que a maioria das opiniões é positiva, mas parte da informação é perdida. Já na Classificação em nível de aspecto, cada uma das características seria analisada.

A oração (2) pode ser dividida em duas sentenças, (2.1) e (2.2). As duas sentenças classificam o objeto como positivo.

Já na oração (3) há opiniões contrárias em torno de um mesmo objeto. Ao analisarmos a sentença (3.1) há uma opinião positiva em relação ao objeto Spettus, porém, na sentença (3.1) há uma opinião negativa em relação ao mesmo objeto. Nas duas sentenças o que difere são os aspectos "carne" na sentença (3.1) e "sushi" na sentença (3.2). O primeiro é classificado positivamente, e o segundo é classificado negativamente. Ou seja, embora o objeto seja o mesmo, os aspectos analisados são diferentes.

Agora serão mostradas cada uma das definições, sendo exemplificadas para deixar mais claro conceito:

- **Opinião**: Uma opinião é uma visão, atitude, sentimento, emoção ou avaliação sobre um aspecto ou objeto por parte de um detentor de opinião. Podendo ela ser positiva ou negativa.

- **Orientação da Opinião:** A orientação da opinião sobre um aspecto ou objeto, pode ser positiva, negativa ou neutra. Ao longo do trabalho, o termo "polaridade" é usado para se referir à orientação da opinião.

- **Objeto:** Um *objeto* é uma entidade que pode ser um produto, pessoa, evento, organização ou tópico. A ela está associado o seguinte par, *o*: (T, A), onde T é uma hierarquia de *componentes* (ou *partes*) e *subcomponentes* e A é um conjunto de atributos. Cada componente possui seu próprio conjunto de subcomponentes e atributos. Ao longo do trabalho, o termo *produto* é usado para se referir a objeto.

- **Aspecto:** Um aspecto é uma característica, atributo, propriedade, parte ou componente de um objeto, seja este um produto, componente, serviço, pessoa, evento ou organização. E cada aspecto pode ser classificado com *explícito*, quando aparece no texto, ou pode ser *implícito*, quando, apesar de não estar no texto, ele puder ser deduzido a partir do contexto. Um exemplo de aspecto implícito pode ser visto no **exemplo 2.1** na sentença (2.2), onde 'caro' qualifica o aspecto 'preço', apesar do aspecto não aparecer no texto.

- **Palavras Opinativas:** São as palavras que qualificam os aspectos. Na maioria das vezes são os adjetivos e advérbios.

Como foi mostrado anteriormente no **exemplo 2.1** na oração (3), apesar de ser o mesmo objeto (Spettus) nas duas sentenças, os aspectos analisados são diferentes (carne e sushi).

A imagem abaixo (Figura 2.1) irá deixar mais clara a distinção entre um objeto e seus aspectos:

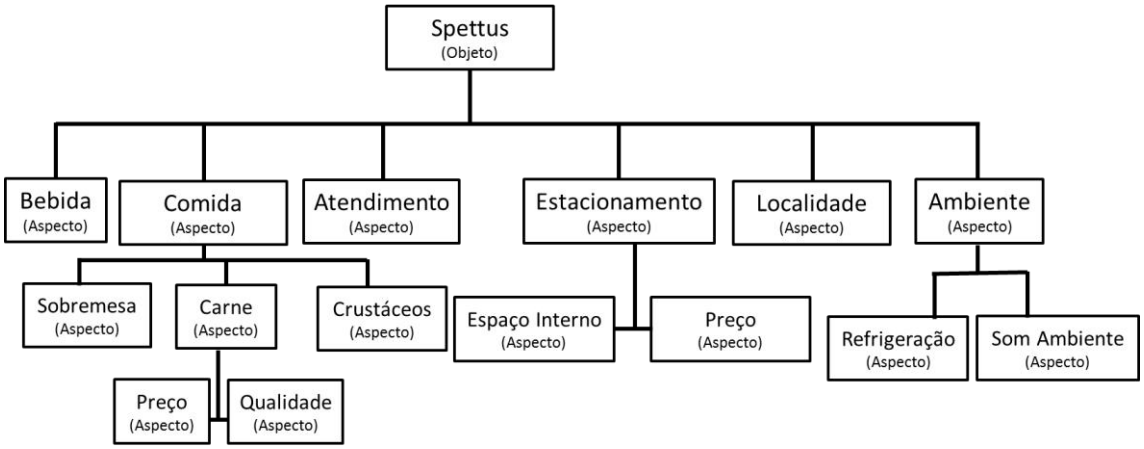

Figura 2.1: ontologia de aspectos do objeto *Spettus*

A figura 2.1 mostra o exemplo de um objeto, e os aspectos relacionados a ele. Em alguns casos, cada aspecto pode ser subdividido em outros dois ou mais aspectos. É o caso do aspecto comida que pode ser subdividido nos aspectos sobremesa, carne e crustáceos.

Outra definição importante para o bom entendimento do trabalho é a dos diferentes tipos de classificação. Segundo [LIU, 2010], a Análise de Sentimentos pode ser tratada em dois níveis distintos, e com objetivos diferentes:

- **Classificação em Nível de Documento:** a opinião está em torno de todo texto ou sentença em questão. A opinião é dada como positiva, negativa ou neutra, observando o texto ou sentença por completo.
- **Classificação Baseada em Aspecto:** No lugar de classificar o texto como um todo, cada aspecto é classificado.

Analisando novamente o **exemplo 2.1**, podemos ver claramente a diferença entre os dois tipos de análises.

Na classificação em nível de texto, será obtida a Orientação da Opinião referente ao texto completo, que pode ser positiva, negativa ou neutra. Porém, diversas informações são perdidas, já que dentro do texto há tanto informações negativas como positivas. Já ao utilizar a Classificação Baseada em Aspecto, a opinião referente a cada um dos aspectos mostrado no texto é considerada. No caso da oração (3), há duas opiniões contrárias para o mesmo objeto, porém para aspectos diferentes. Enquanto na Classificação em Nível de Texto essa informação é perdida, na Classificação Baseada em Aspecto ela é mantida.

Porém, nem todas as sentenças são classificáveis, então é importante antes de classificar separar as sentenças classificáveis, das não classificáveis [LIU, 2006]. Abaixo serão mostrados os termos e as definições usadas para esses dois tipos de sentenças:

- **Sentença Objetiva:** A sentença não apresenta a opinião do autor, são mostradas apenas informações factuais sobre o objeto em questão.
- **Sentença Subjetiva:** A sentença apresenta opinião ou crença do autor em relação ao objeto em questão.

Exemplos claros dessas duas definições podem ser vistas no **exemplo 2.1**. Enquanto a oração (1) apresenta uma Sentença Objetiva, as orações (2) e (3) apresentam sentenças Subjetivas.

Continuando a apresentação de conceitos, podemos ver mais definições:

- **Detentor da Opinião:** É o autor, pessoa ou organização, que expressa a opinião. Na maioria das ferramentas de expressão, como blogs e fóruns. O detentor da opinião são os responsáveis por postar os comentários. E em alguns casos, a origem da opinião não é do autor do comentário, porém, o Detentor da Opinião é referenciado no texto. Por exemplo, em "Meu irmão não gostou da comida do Spettus.", nesse caso, o Detentor da Opinião é o irmão do autor do texto.

- **Opinião Direta:** É a opinião que faz referência a apenas um objeto ou aspecto.

- **Opinião Comparativa:** É a opinião que apresenta uma relação comparativa das semelhanças ou diferenças entre dois ou mais objetos. Uma opinião comparativa é normalmente expressa usando um advérbio de comparação [LIU, 2010].

Exemplos de opiniões diretas estão em **exemplo 2.1**. Já um exemplo de opinião direta pode ser a seguinte sentença: "A carne do Spettus é tão boa quanto à do Boi Preto.".

A opinião comparativa é mais complexa de ser tratada, pois ela exige estudos mais aprofundados [MAGALHÃES, 2009]. Neste trabalho serão tratadas apenas sentenças com opinião direta.

2.1.1. Pointwise Mutual Information

Pointwise Mutal Information (PMI) é uma medida de Teoria da Informação que avalia a relação entre uma ou mais palavras em um texto, comparando a probabilidade de encontrar duas palavras juntas com a probabilidade de encontrá-las separadas, de modo a informa qual a associação entre elas [THOMAS & COVER, 1991].

Para encontrar o PMI de das palavras *ent*e *pal*, com probabilidades de aparecerem no texto P(*ent*) e P(*pal*), dar-se da seguinte forma:

$$PMI(ent, pal) = \log\left(\frac{(P(ent \wedge pal))}{P(ent)P(pal)}\right) \qquad \text{(Equação 2.1)}$$

Onde $P(ent \wedge pal)$ é a probabilidade das palavras *ent*e *pal* aparecerem próximas. A relação de proximidade pode ser considerada uma distância *d* previamente determinada, de aparecerem na mesma sentença ou no mesmo documento.

Dessa forma, através do PMI é possível informar o quanto duas palavras estão relacionadas entre si. Em alguns trabalhos [LOPES et al., 2008] [FERNANDES, 2010], a medida é usada para verificar a relação de algumas palavras com uma classe de palavras positivas ou negativas, avaliando assim, a polaridade da palavra em questão.

2.2. Algumas Aplicações

Apesar de uma área bastante recente, existem diversas aplicações para uso da Análise de Sentimentos, algumas já no mercado, e outras em trabalhos de pesquisa. Abaixo estão listadas algumas dessas aplicações:

- **Análise de Empresas na Bolsa de Valores:** A Vetta Labs lançou na TechCrunch 50 em 2008 a O StockMood.com, uma ferramenta para auxiliar pequenos investidores na bolsa dos EUA. Ela identifica o humor do mercado em relação às empresas negociadas na bolsa de valores baseado nas opiniões dos analistas, com o objetivo de identificar a tendência dos preços da Bolsa de Valores.

- **Análise de Um Produto:** Essa é um das apliações mais comuns para AS. Opinião dos usuários torna-se um fator de decisão na hora da compra de um produto, ou até mesmo para melhoras nos produtos das empresas. Um exemplo de aplicação com esse intuito é *Sentweet* da empresa Vetta Labs e o Twittersentiment[1] usados para classficar as opiniões postadas no microblog Twitter.

- **Análise de Políticos:** Eleitorando[2] é um software com o objetivo de indentificar as opiniões dos usuários do Twitter e do Youtube em torno dos polícos. O software analisa as opiniões dos usuários disponibilisa as informações através de gráficos para o cliente.

- **Outras Aplicações:** opSys[3] é um sistema de Mineração de Opinião para conteúdo *On line*, que indica a orientação semântica dos artigos filtrados, traçando um panorama de quanto as entidades pesquisadas estão sendo citadas positivamente ou negativamente.

2.3. Etapas da Análise de Sentimentos

Com o objetivo de atingir melhores resultados, a análise da opinião de um texto é dividida em tarefas que normalmente são sequenciais e complementares. Tamanha é a complexidade de cada uma dessas tarefas que, geralmente, os trabalhos focam em uma tarefa ou duas tarefas, deixando o restante para ser realizado com a intervenção do usuário [SIQUEIRA, 2010].

[1] http://twittersentiment.appspot.com/
[2] http://www.eleitorando.com.br
[3] http://www.opsys.com.br/

Nas próximas seções serão descritas essas tarefas, que são: análise da subjetividade do texto, extração das características, classificação da opinião, visualização e sumarização.

Para analisar algumas dessas fases, será utilizado o seguinte texto:

> *"Se eu você for para o Spettus, eu irei ao Boi Preto. A comida do Boi Preto é maravilhosa, e a sobremesa servida no Boi Preto não é cara. O sushi no Spettus é muito bom, mas dizer que o camarão é bom é piada."*

Exemplo 2.2: texto criado pelo autor.

2.3.1. Subjetividade

Como foi mostrado na seção de Conceitos Básicos, Objetividade está relacionado a informações factuais, enquanto Subjetividade está relacionado a opiniões ou crenças do autor. Analisando esses dois conceitos, é fácil verificar que elas representam definições contrárias.

A primeira tarefa a ser executada é a identificação da opinião e detecção de subjetividade, separando as Sentenças Objetivas das Sentenças Subjetivas. As únicas sentenças que terão utilidades para AS são as subjetivas, pois elas apresentam um sentimento, emoção ou pensamento relativo a algo, ou seja, representam as opiniões que serão analisadas.

Essa tarefa, por muitas vezes, se torna mais difícil que realizar a própria classificação da polaridade do texto. Logo, a técnica usada na Identificação da Opinião poderá impactar nos resultados da classificação do sentimento [PANG & LEE, 2008].

Utilizando-se do exemplo do **exemplo 2.2**, citado anteriormente, temos três sentenças das quais a primeira não identifica uma opinião, enquanto as outras duas mostram duas opiniões claras em relação aos objetos Spettus.

Uma das técnicas usadas para identificar a subjetividade das sentenças é proposta por Hatzivassiloglou e Wiebe em [HATZIVASSILOGLOU & WIEBE, 2000], que faz a análise da subjetividade das sentenças através das orientações dos adjetivos. Outras técnicas podem ser observadas em [WIEBE et al., 2004] [WILSON et al., 2004] [WIEBE & MIHALCEA, 2006] [OUNIS et al., 2006].

2.3.2. Extração de Características

Após identificar todas as sentenças que expressam opiniões, o próximo passo é responsável por identificar os aspectos ou características que estão ligados ao objeto a ser analisado. Ao se ter uma análise positiva em relação a um produto, não significa que todos os aspectos foram positivamente classificados em relação ao objeto analisado, porém, a maioria das características foi classificada positivamente, podendo algumas das características serem classificadas negativamente.

Verificando o exemplo 2.2 os aspectos extraídos ali seriam: comida, carne e sushi, pois esses revelam características que podem ser analisadas quando comparados dois ou mais restaurantes. Esses são os aspectos a serem classificados no texto.

A extração de características é uma das tarefas mais difíceis de ser realizada, e menos propensa a automação da Análise de Sentimento. Em muitos casos é preciso considerar o domínio, ou seja, utilizar uma técnica específica para um determinado domínio proposto, para que seja possível extrair todas as características sem cometer erros [SIQUEIRA, 2010].

Para [MAGALHÃES, 2008], os sistemas de Extração de Informação apresentam algumas fases, apesar de existirem muitas variações de sistema para sistema:

- O texto é inicialmente dividido em sentenças e palavras, para a aplicação de um Pos-Tagger, que irá aplicar etiquetas para cada uma de suas classes gramaticais;
- A próxima etapa é a análise da sentença que compreende uma fase de identificação dos grupos de substantivos, dos grupos de verbos, das expressões preposicionais e das outras estruturas simples.
- A próxima fase é específica para o domínio de aplicação, pois essa é a etapa que identifica as entidades relevantes no texto, ou seja, identifica os aspectos a partir do grupo de susbstantivos anteriormente selecionados.
- A próxima fase é de "merging", onde acontece a resolução co-referenciada ou resolução anafórica: o sistema examina cada entidade encontrada no texto e determina se tal entidade se refere a uma entidade já existente ou se ela é nova e deve ser adicionada ao nível de discurso do sistema que representa o texto.

Existem muitas técnicas para a extração de aspectos. Existem as técnicas que usam as abordagens baseadas em estatística e lingüística, quando se utiliza da classe gramatical, já que a maioria das características ou aspectos são os substantivos das sentenças, e a frequência com que elas aparecem nas sentenças que tratam um determinado assunto [HU & LIU, 2004].

O próximo passo da AS é a classificação desses aspectos extraídos durante a fase de extração em cada sentença. A próxima seção trata dessa classificação, que determina a polaridade das características, se positiva, negativa ou neutra.

2.3.3. Classificação

Depois de identificar as características referentes ao domínio, o próximo passo será classificar a polaridade da sentença. Essa classificação pode ser dividida em níveis, e diversas técnicas diferentes podem ser aplicadas para chegar a um melhor resultado da classificação.

A classificação pode ser dividida em níveis, ou seja, pode ser classifica em nível de aspecto, que é o nível de maior granularidade dos documentos, quando a preocupação maior está em classificar cada característica do documento. E a classificação ainda pode ser feita em nível de sentença ou documento [FERNANDES, 2010].

Para realizar a classificação, vários passos são realizados, utilizando-se de diversas técnicas. Abaixo são descritos os passos de um algoritmo de classificação [DING et al., 2008]:

- **Encontrar e classificar as palavras opinativas**: Esse primeiro passo encontra as palavras opinativas contidas na sentença e classifica-as como positivas, negativas ou neutras. Como a primeira sentença do exemplo 2.2 não representa uma opinião, então iremos utilizar apenas as duas ultimas sentenças. As sentenças podem ser

classificadas como, *A comida do Boi Preto é maravilhosa **[+1]**, e a sobremesa servida no Boi Preto não é cara **[-1]**. O sushi no Spettus é muito bom **[+1]**, mas dizer que o camarão é bom **[+1]** é piada.".*

- **Cláusulas Negativas:** Um segundo passo é identificar as cláusulas negativas, que irão inverter a polaridade das palavras opinativas as quais ela se referem. Na segunda sentença do exemplo 2.2, apesar de a segunda palavra opinativa ser classifica como negativa, a sua polaridade é invertida por causa da cláusula negativa que a antecede. Logo, a sua classificação iria ficar da seguinte forma: "*A comida do Boi Preto é maravilhosa **[+1]**, e a sobremesa servida no Boi Preto não é cara **[+1]**".*

- **Cláusulas Adversativas:** Palavras que tratam de oposição, como as cláusulas adversativas, mostram opiniões contrárias em relação ao mesmo objeto em estudo. A ultima sentença do exemplo 2.2 mostra um exemplo: *"O sushi no Spettus é muito bom **[+1]**, mas dizer que o camarão é bom **[+1]** é piada.".* Nesse caso, apesar da ultima palavra opinativa ser classificada como positiva, por causa da conjunção "mas", as duas palavra opinativas, devem conter polaridades contrárias. Logo, a polaridade da sentença fica da seguinte forma: *"O sushi no Spettus é muito bom **[+1]**, mas dizer que o camarão é bom **[-1]** é piada.".*

Em uma pesquisa recente [LIU et al., 2009] , foi visto que as sentenças condicionais podem mostrar algumas particularidades ao serem classificadas como positivas, negativas ou neutras. Outras abordagens de algoritmos podem ser vistos em [LIU, 2010] [POPESCU & ETZIONI, 2005].

2.3.4. Visualização e Sumarização

Esta ultima tarefa é responsável por agregar e representar os resultados da AS. As informações obtidas a partir da classificação das sentenças podem ser mostradas de diversas formas, desde textos descrevendo os resultados da análise a gráficos que fazem comparações entre dois ou mais produtos do mesmo domínio, por exemplo, um gráfico comparar diversos aspectos em relação aos objetos "Spettus" e "Boi Preto", no domínio de restaurante.

Bo Pang em [PANG & LEE, 2008], divide a Visualização e Sumarização em dois tipos distintos: Sumarização de documento simples e a Sumarização de documentos com diversos objetos:

- **Sumarização de Documentos Simples:** Na sumarização de textos que tratam de um único objeto, na sumarização, algumas abordagens realizam a sumarização baseadas na extração de sentenças ou textos similares. O objetivo é unir os textos que relatam de opiniões parecidas ou que relatam de um mesmo aspecto, para mostrar os pontos positivos e negativos de cada aspecto.

- **Sumarização de Multi-Documentos:** Este tipo de sumarização está relacionado ao resultado de vários objetos (ou produtos) analisados. O intuito dessa análise não é só mostrar ao usuário os pontos positivos e negativos de cada aspecto, mas também comparar cada um deles. Abaixo é mostrado um exemplo da sumarização de um resultado da comparação entre três celulares de marcas diferentes:

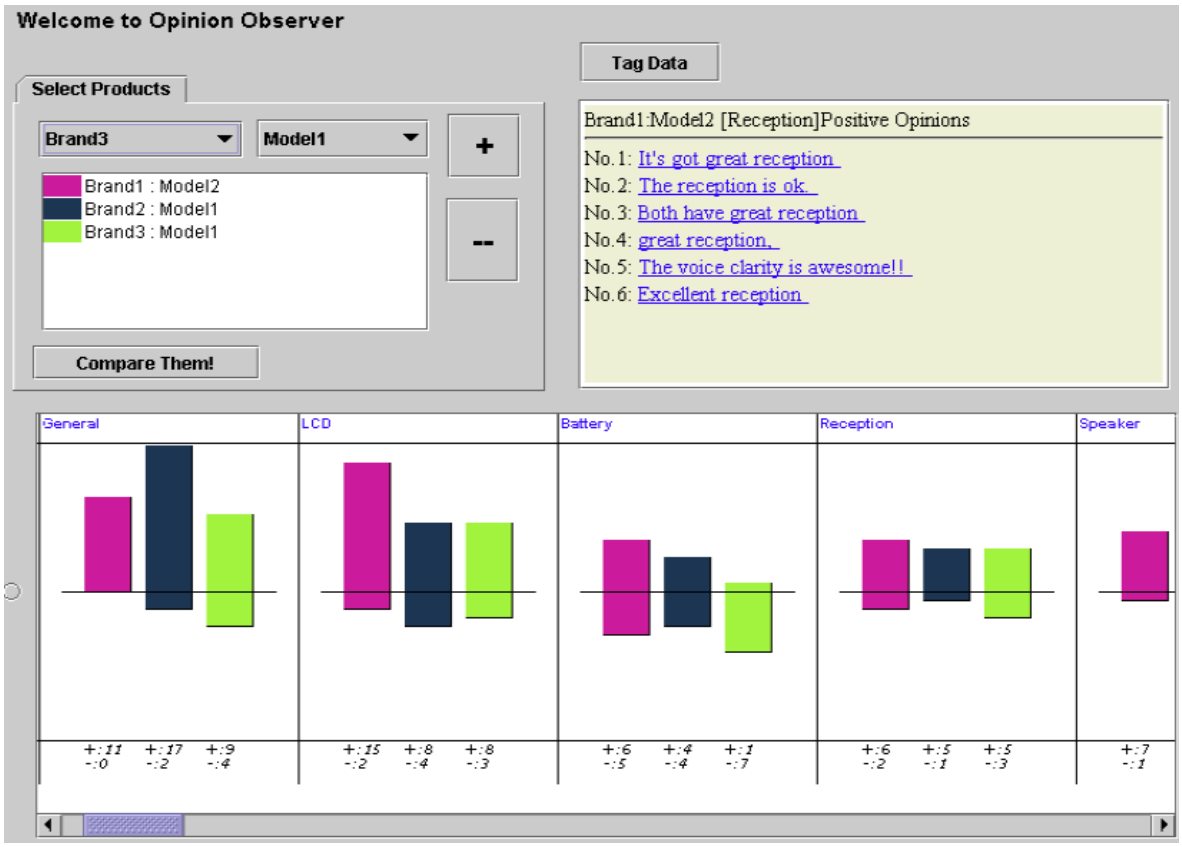

Figura 2.2: Três celulares foram representados, cada um com sua cor. Para cada aspecto (LCD, Bateria, Recepção, etc), três barras são mostradas, cada uma das barras representa um celular. Quanto mais acima da linha horizontal a barra estiver, mais avaliada positivamente ela foi. Quanto mais abaixo da linha horizontal, mais negativa foi a avaliação [LIU et al., 2005].

Mostradas as etapas para a AS, na próxima seção serão mostrados os principais desafios e limitações da área.

2.4. Desafios e Limitações da Área

Análise de Sentimentos é uma subárea de Processamento de Linguagem Natural bastante recente, o que faz com que ainda existam muitos desafios e limitações a serem superados. Nas próximas seções, serão citados alguns dos desafios enfrentados na AS.

2.4.1. Identificação da Subjetividade

Distinguir se um texto é uma opinião ou um fato ainda é um grande desafio em AS. Para Mihalcea em seu trabalho [MIHALCEA et al., 2007], identificar se um texto é opinativo é, geralmente, mais complexo que a própria classificação do texto como positivo ou negativo.

Hatzivassiloglou e Wiebeem em seu trabalho [HATZIVASSILOGLOU & WIEBE, 2000] estudaram os efeitos dos adjetivos em uma sentença. Eles identificaram que a partir da analise dos adjetivos é possível afirmar a subjetividade das sentenças.

Apesar de existirem tantos estudos na área, a subjetividade ainda é uma das tarefas mais complexas da AS.

2.4.2. Classificação da Polaridade dos Adjetivos e dos Advérbios

Classificar a polaridade de um adjetivo não é apenas uma tarefa complexa, como também com poucos resultados na literatura.

Um dos pontos iniciais para se classificar um texto é a classificação dos adjetivos e advérbios. Em muitos trabalhos, essa classificação é feita pelo próprio usuário [FERNANDES, 2010], ou a partir de bases da internet, que na maioria das vezes não consideram o domínio da aplicação [ESULI & SEBASTIANI, 2006].

Entraremos em mais detalhes de algumas soluções encontradas, e os problemas que podem acarretar essa etapa da AS no próximo capítulo.

2.4.3. Outros Desafios com Processamento de Texto

Diversos desafios são encontrados ao processar o texto com a opinião. O primeiro deles é identificar textos com ironias e sarcasmos, pois esses tipos de sentenças podem levar a uma classificação errônea do texto. Em [CARVALHO et al., 2009], eles criam algumas "regras" para se detectar as ironias nos textos, analisando formas diminutivas, interjeições, morfologia do verbo, entre outros. Apesar de eles terem mostrado ser possível detectar a ironia nos textos, ainda apresenta uma taxa de precisão baixa.

O uso de pronomes para referenciar itens também pode dificultar a análise das sentenças. E como o uso de pronomes é bastante comum em qualquer tipo de texto, esse fator é um dos grandes desafios enfrentados pela AS .

Palavras erradas e jargões da internet como, "vc", "tb", "doro", "naum", ":)", têm sido tão comuns no meio de internautas, que alguns vocábulos criados por eles influenciam a maneira de escrever e de falar, muitas vezes originando novas linguagens [FARIA & ZUQUIM, 2005]. Pelo fato da identificação desses jargões não ser uma tarefa trivial, pois existem grande quantidade deles que mudam de acordo com a cultura local e país, identifica-los não é uma tarefa trivial, o que pode piorar o resultado final da AS.

2.5. Conclusão

A AS é uma processo bastante complexo, dividida em diversas etapas, podendo uma comprometer nos resultados da etapa posterior, por isso, em muitos trabalhos, algumas dessas etapas são realizadas de forma manual, sendo o foco do trabalho, apenas uma ou duas dessas tarefas.

Por essa área ser bastante recente, ainda existem muitos desafios para serem superados. E como foi dito pelo **GartnerGroup**, a AS é uma das grandes tendências para o futuro, então é de se esperar um grande crescimento dessa área.

No próximo capítulo mostraremos, com mais alguns detalhes, o processo de Classificação de Sentimentos por este ser o principal foco deste trabalho.

3. Classificação de Sentimentos

Nesse capítulo será mostrada com mais detalhes a fase de classificação da AS, por ser o principal foco deste trabalho. Na primeira seção deste capítulo, será mostrada uma breve motivação para o estudo da classificação da AS. A seguir, será mostrado o trabalho proposto em [FERNANDES, 2010], que serviu de base para este TG; na seção 3.3 será feita uma breve descrição do SentiWordNet [ESULI & SEBASTIANI, 2006], abordagem usada para melhorar o processo proposto em [FERNANDES, 2010]; e por fim, serão feitas algumas considerações finais.

3.1. Motivação para trabalhar com classificação

A fase de classificação da AS é uma das principais, pois é nessa etapa que ocorrerá a identificação da polaridade do texto, que é o principal objetivo da AS.

Diversas técnicas são usadas para classificar um texto, desde algoritmos de aprendizagem de máquinas, que é a abordagem dominante para resolver este problema [SEBASTIANI, 2002], a abordagens estatísticas [LOPES et al., 2008], que usaremos no trabalho em questão.

3.1.1. Desafios e Limitações

Os algoritmos de aprendizagem de máquina possuem bons resultados, porém muita informação é perdida, pois a partir deles, apenas é possível realizar a classificação ao nível de documento, ao contrário das abordagens estatísticas.

Contudo, os métodos estatísticos ainda são pouco usados nesta área, pois esses métodos exigem uma prévia classificação das palavras opinativas. Essa classificação é geralmente feita pelo usuário, tornando o processo, muitas vezes, impraticável devido à grande quantidade de palavras opinativas que podem ser encontradas em um corpus de texto ou em um determinado domínio.

[TURNEY, 2002] apresenta um algoritmo de classificação não supervisionado. O processo proposto por Turney faz a classificação nas seguintes etapas: primeiro utiliza-se de um Por-Tagger para determinar a classe gramatical de cada palavra; logo após é feito a classificação dos adjetivos e verbos através do algoritmo PMI-IR [TURNEY, 2002]. O cálculo da polaridade das palavras é feito a partir da diferença entre sua similaridade com a palavra positiva e a sua similaridade com a palavra negativa. Embora o algoritmo tenha atingido uma precisão média de 74%, ainda é preciso de uma classificação manual de algumas palavras previamente.

Em outro trabalho, Pang [PANG et al., 2002] propõe um algoritmo de aprendizagem automática para classificar corpora de críticas de cinemas. Dos diversos algoritmos de aprendizagem automática, o que se saiu mais eficiente foi o Support Vector Machine (SVM), atingindo uma taxa de precisão de 82,9%. Porém, como nos trabalhos [LIU, 2010] [TURNEY, 2002] [SANTORINI, 1995] [KOTSIANTIS et al., 2007] [PANG & LEE, 2008] [LIU, 2006] [MAUÁ, 2009], a classificação é feita em nível de documento, perdendo importantes informações do texto.

3.2. O Sistema HowGood

Nesta seção será mostrado o sistema proposto por [FERNANDES, 2010], trabalho que serviu de base para esse Trabalho de Graduação. Nas duas próximas seções será mostrada uma visão geral do HowGood e algumas deficiências do sistema.

3.2.1. Visão geral

Como foi dito anteriormente a classificação pode ser feita em nível de documento ou em nível de aspecto. O método utilizado pelo HowGood trata a análise em nível de aspecto que fornece uma maior granularidade no detalhamento dos resultados. Dessa forma obtêm-se informações além da polaridade do texto, nesse caso é analisado cada atributo, ou aspecto.

O seu processo é dividido em cinco etapas, que vão desde a coleta de dados na rede social Twitter a determinação das polaridades dos aspectos. As cinco etapas são as seguintes:

- **Monitoramento de menções a marcas no Twitter:** Nessa etapa o usuário entra com a marca que deseja analisar no sistema, que faz uma busca através da API do Twitter [TWITTER, 2010]. Nesse passo o sistema roda rotinas que capturam e armazenam na base de dados as postagens do Twitter que contêm a marca alvo a ser analisadas.

- **Etiquetagem de documentos:** Nessa etapa do sistema é realizada a etiquetagem das palavras. Inicialmente as postagens são processadas e submetidas ao Tree-Tagger [GAMALLO, 2005], um Pos-Tagger que realiza a etiquetagem da classe gramatical para diversos idiomas, inclusive para o português. No caso do sistema, cada palavra é classificada segundo a sua classe gramatical para o idioma português, mas, segundo o autor, pelo fato do sistema está bem modularizado, qualquer idioma pode ser utilizado, sendo necessário apenas modificar o Pos-Tagger utilizado.

- **Modelagem do Domínio:** Nessa etapa do sistema, são determinados os aspectos e as palavras opinativas que se encontram em cada uma das postagens. Essa tarefa é realizada de forma semiautomática, ou seja, com interferência do usuário na maior parte do processo. Logo após a determinação da classe gramatical de cada palavra na etapa anterior, são criados três conjuntos: o dos nomes (NOM), o dos adjetivos (ADJ) e o dos advérbios (ADV); os outros termos são desconsiderados. Para determinação dos aspectos, os nomes (NOM), são ordenados em ordem decrescente de frequência e mostrados na tela, para que o usuário possa selecionar os termos correspondentes aos aspectos, como mostra a figura abaixo:

[1] http:// twitter.com

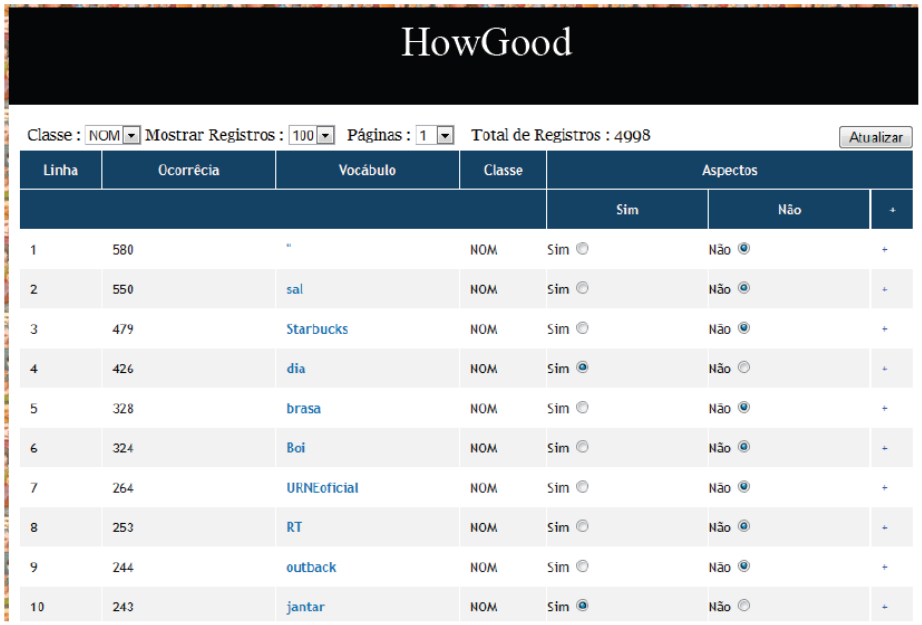

Figura 3.1: Tabela de relevância de aspectos [FERNANDES, 2010]

Na próxima etapa, similar a anterior, serão selecionados as palavras opinativas do texto, e atribuída as suas polaridades. Essas palavras darão o sentido positivo ou negativo dos aspectos. As palavras opinativas são compostas pelos adjetivos e advérbios do texto.

Inicialmente, são separados os adjetivos (ADJ) e advérbios (ADV), que são postos em ordem decrescente, segundo a sua frequência em relação ao corpus de documentos usado na modelagem do domínio.

Para determinar a polaridade de cada uma das palavras opinativas, a listagem é mostarda ao usuário, que irá determinar a polaridade de cada um dos termos como positivo, negativo ou neutro, como é mostrado na figura 3.2 e na figura 3.3.

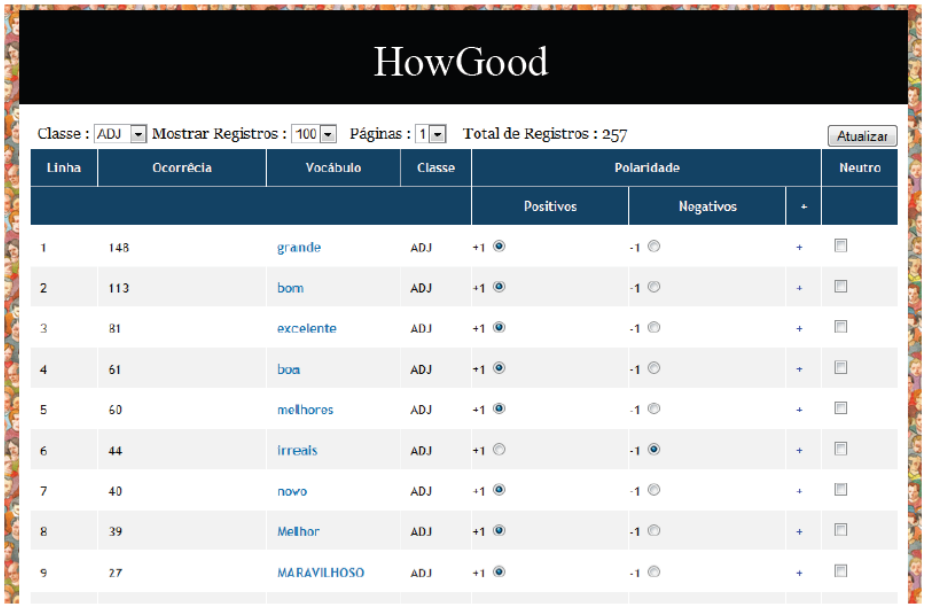

Figura 3.2: Tabela de relevância de palavras opinativas como adjetivos [FERNANDES, 2010]

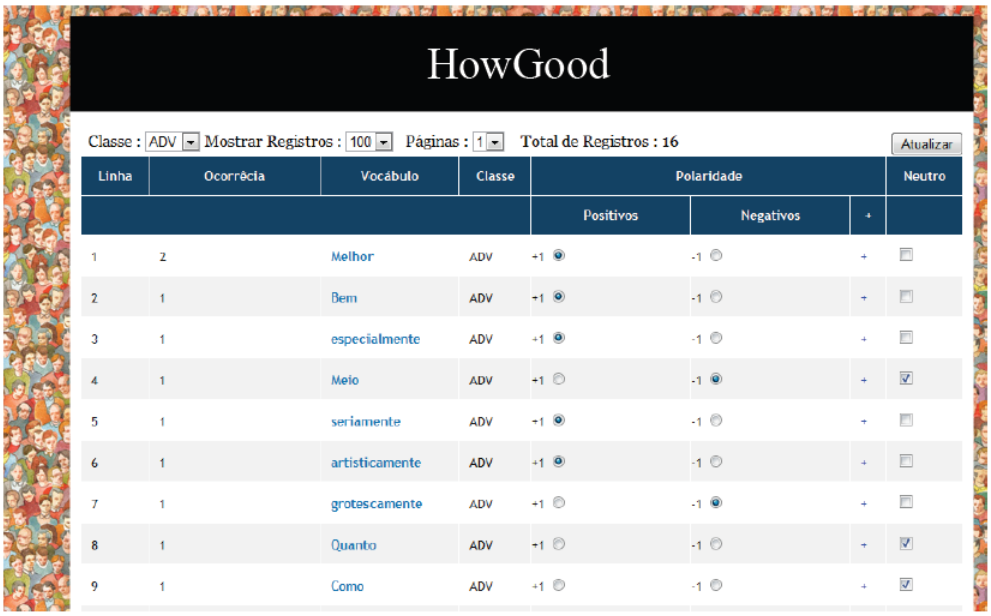

Figura 3.3: Tabela de relevância de palavras opinativas como advérbios [FERNANDES, 2010]

O último passo da modelagem do domínio trata da definição do espectro de influência das palavras opinativas. Espectro de influência consiste na distância média na qual cada aspecto se deixará influenciar por uma palavra opinativa. Ou seja, a partir do espectro de influência, será possível saber quais termos devem ser considerados no cálculo do sentimento associado a cada um dos aspectos. Por exemplo, no texto "A carne é macia e saborosa, mas é cara.". Se o espectro de influencia do aspecto *carne* for quatro, então os termos *macia* e *saborosa* irão entrar no cálculo do sentimento do aspecto, porém o termo *cara* não vai entrar por estar a uma distância maior que quatro do aspecto.

- **Determinação da Orientação da Opinião:** Esta etapa é responsável por classificar os aspectos que foram selecionados pelo usuário na etapa anterior. Essa etapa é divido em três passos: o primeiro é atribuir a polaridade a cada uma das palavras opinativas; o segundo é tratar as expressões negativas; e o terceiro é ligar a palavra opinativa ao aspecto. A partir do exemplo abaixo faremos os três passos.

''O Spettus é ótimo e a comida não é cara.''

No exemplo acima o primeiro passo é atribuir polaridade a cada uma das palavras opinativas, ficando da seguinte forma:

"**O Spettus é ótimo (+1) e a comida **não** é cara (**-1**).**"

O segundo passo irá tratar as expressões negativas, pois elas irão inverter a polaridade dos termos. As expressões negativas invertem a primeira palavra opinativa à direita. Ou seja, o texto ficará da seguinte forma:

"O Spettus é ótimo (+1) e a comida* não *é cara (+1)."

O último passo é a ligação entre as palavras opinativas e os seus respectivos aspectos. Nesse passo, é usado o espectro de influência calculado na etapa anterior. Logo, se o espectro de influência tiver medida 3 para os dois aspectos encontrados no texto, *Spettus* e *comida*, então teremos as seguintes polaridades para os aspectos: *Spettus* com a polaridade de *ótimo* (+1) e *comida,* com a polaridade do termo *cara* (+1).

- **Apresentação dos Resultados da Análise de Sentimentos:** Nessa etapa do processo, são mostrados os resultados obtidos de duas formas diferentes: a sumarização por marca e a sumarização por comentário. No primeiro caso, tem-se uma visão geral de todas as marcas sendo monitoradas com a porcentagem dos comentários positivos, negativos e neutros. Já na sumarização por comentário, são listados os comentários com a análise de sentimento realizado em cada um dos documentos.

3.2.2. Crítica

Apesar de bem estruturado e modularizado, o sistema proposto encontra algumas deficiências, como na escolha da base e na ausência de algumas técnicas que poderiam melhorar a classificação. Porém, a principal deficiência está na não trivial tarefa da classificação das palavras opinativas pelo usuário.

Os comentários que o sistema usa para fazer as análises são retirados da base de dados do Twitter, ou seja, cada comentário possui no máximo 140 caracteres. Porém, textos curtos como a do Twitter, são geralmente usados para a classificação em nível de documento, pois nessas aplicações o usuário está apenas interessado em saber a quantidade de avaliações positivas e negativas [SIQUEIRA, 2010]. E pelo fato dos textos serem curtos é comum a ausência de diversas palavras e de abreviações de outras, prejudicando assim o processo de classificação por aspectos.

O autor dá a seguinte explicação para não considerar as cláusulas adversativas:

"Cláusulas adversativas não foram consideradas, pois toda e qualquer palavra opinativa já carrega sua própria polaridade sem a necessidade de se inferir por tais cláusulas." [FERNANDES, 2010]

Porém, essas expressões podem mudar a polaridade das opiniões, e até mesmo de alguns aspectos, principalmente quando tratamos de textos com ironia. Ao analisarmos a seguinte frase podemos verificar esse fato:

"A carne pode até ser saborosa, mas dizer que o ambiente é agradável é demais."

Ao analisarmos o texto acima sem considerar a cláusula adversativa *'mas'*, a polaridade do aspecto *ambiente* será positiva, o que não é a realidade. Alguns trabalhos [KAJI & KITSUREGAWA, 2007] [WIEBE & WILSON, 2002] consideram as técnicas adversativas, pois elas trazem a informação de sentimento oposto entre as sentenças que se encontram antes e depois da conjunção. Nesse exemplo, apesar do termo *agradável* relatar uma qualidade positiva para o aspecto *ambiente*, não é o real sentimento do autor do comentário, que descreve um texto com ironia. Logo, se analisarmos o exemplo acima considerando a cláusula adversativa, teríamos a seguinte classificação:

*"A carne**(+1)**pode até ser saborosa, mas dizer que o ambiente**(-1)** é agradável é demais."*

O aspecto *Spettus* recebe a polaridade positiva, pois saborosa relata uma qualidade positiva. Porém, *ambiente* recebe polaridade negativa, pois, apesar de *agradável* relatar uma polaridade positiva, a conjunção adversativa tende a colocar uma polaridade contrária a da sentença anterior à conjunção, ficando o aspecto *ambiente* com polaridade negativa. Ou seja, quando há cláusulas adversativas, é verificado se as polaridades são opostas, caso possuam a mesma polaridade, uma delas é invertida.

Outro ponto a se considerar no HowGood é a tarefa de classificar as palavras opinativas. Essa é uma tarefa atribuída para o usuário, porém, se analisarmos o contexto Web, se torna não trivial. Na Web, a grande quantidade de textos que podem ser analisados, e a grande quantidade de palavras opinativas (adjetivos e advérbios), que podem ser geradas, tornam uma classificação manual dessas palavras uma tarefa não praticável. Por isso, automatizar essa tarefa é fundamental para tornar esse sistema usável.

3.3. SentiWordNet

As pesquisas em Análise de Sentimentos têm crescido bastante nos últimos anos. Pesquisas recentes têm buscado determinar automaticamete a polaridade de termos subjetivos, ou seja, determinar se um termo que é um marcador de conteúdo opinativo (palavras opinativas, como adjetivos e adverbios) tem um efeito positivo ou uma conotação negativa. Um trabalho recente que tem seguido esse ramo de pesquisa é o SentiWordNet [LOPES et al., 2008], que é um recurso lexical em que cada palavra está associada a três escores numéricos Obj, Pos e Neg, descrevendo a conotação neutra, positiva e negativa da palavra, respectivamente.

SentiWordNet (SWN) é livremente diponível para fins de pesquisa, e é dotado de uma interface gráfica baseada na web. A base do SWN é um arquivo de texto formado por mais de 110 mil palavras inglesas, divididas entre verbos, substantivos, adjetivos e advérbios. Para cada uma das palavras é associado um score positivo (Pos) e um score negativo (Neg), a diferença de 1 com a soma desses dois scores dá o score neutro (Obj) – ver Equação 3.1.

$$\mathbf{Obj = 1 - (Pos+Neg)} \qquad \text{(Equação 3.1)}$$

Cada palavra é seguida de um exemplo de seu uso em inglês. Cada linha do arquivo de texto vem com as informações referentes a uma palavra. O arquivo está no seguinte formato (exemplo 3.2):

POS	*ID*	*PosScore*	*NegScore*	*SynsetTerms*	*Gloss*
a	00064787	0.625	0	good#5	"the experience was good for her"

Exemplo 3.2: texto tirado da base de dados do SentiWordNet

No exemplo acima, a letra *a* representa a classe gramátical a qual pertence a palavra, no caso cima, *adjetivo*. O número seguinte é um identificador da palavra. Os dois próximos valores, 0.625 e 0 representam o score positivo e o score negativo da palavra, respectivamente. O elemento seguinte é a palavra em questão, no exemplo acima, *good*. E o ultimo elemento representa uma frase de uso da palavra.

Diversos trabalhos utilizam o SWN para identificar a polaridade das palavras. Um deles é [MUSAKAMI et al., 2009], que usa o SWN para capturar os sentimentos das palavras no desenvolvimento do Statement Map, sistema que ajuda usuários a buscar e analisar opiniões na internet nas línguas japonesa e inglesa.

3.4. Considerações finais

Neste capítulo mostramos os principais trabalhos, algoritmos e métodos usados para a etapa de classificação de sentimentos em Análise de Sentimentos, com as principais limitações da área.

Foi apresentado mais detalhadamente o HowGood, trabalho o qual foi baseado esse trabalho, dando uma visão geral do método usado, e suas principais deficiências. E logo após foi mostrado o SentiWordNet um recurso lexical de classificação, que será usado por este trabalho para deixar mais autônoma a tarefa de classificação das palavras opinativas.

No próximo capítulo, será apresentado o sistema que foi o resultado desse trabalho de graduação.

4. BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião

A realidade atual da Web vem nos mostrar uma nova necessidade, a busca por sistemas que possam tratar as opiniões dos usuários nas ferramentas de expressão de opinião da Web, como fóruns, blogs e redes sociais. No capítulo 1 foi mostrado como essa necessidade foi crescendo cada vez mais e junto a ela as pesquisas na área de Processamento de Linguagem Natural, e em particular uma a Análise de Sentimentos, uma subárea de PLN que trata da análise das opiniões encontradas na Web. No capítulo 2 foi mostrada uma breve revisão da literatura em relação a AS, com as principais pesquisas e desafios que ainda enfrentam essa tão recente área. E no capítulo 3 vimos com mais detalhes uma das tarefas da Análise de Sentimento, a Classificação, por se tratar do principal foco dessa pesquisa.

Esse Trabalho de Graduação propõe um sistema que oferece uma solução para essa necessidade da Web, o *BestChoice*, que é um sistema modular para analisar e classificar os textos na Web de forma autônoma.

Nesse capítulo será mostrado todo processo para o desenvolvimento do sistema BestChoice. Na seção 4.1 será mostrada a caracterização do problema, seguido da arquitetura do sistema na seção 4.2. Na seção 4.3 será descrito como foi adquirida a base de dados para análise, e o motivo dessa escolha. Das seções 4.4 à 4.8, serão mostradas todas as etapas do processo de desenvolvimento do sistema.

4.1. Caracterização do Problema

Como visto, um processo de Análise de Sentimento completo é bastante complexo e composto de várias etapas. Desse modo, nem todas as suas etapas foram implementadas no sistema BestChoice.

Neste trabalho de graduação, o foco principal está na fase de Classificação das opiniões. As etapas de Extração e Sumarização foram implementadas, porém de forma simples.

A primeira fase da AS de sentimento, responsável por analisar a Subjetividade do texto, não foi automatizada, considerando-se assim todos os textos de entrada como opinativos. Porém, existem alguns trabalhos na literatura que tratam da etapa de análise da subjetividade do texto, que pode ser acoplada ao sistema de forma simples.

A fase de classificação é uma das principais etapas da AS, por se tratar da fase onde é realizada a caracterização da polaridade das opiniões, e é também a mais complexa, pela necessidade do tratamento dos mais diversos textos (como por exemplo, texto com ironias) e tratar diversos domínios, já que um adjetivo pode classificar positivamente em um domínio e negativamente para outro domínio.

4.2. Arquitetura Geral

A figura 4.1 mostra a arquitetura geral do sistema *BestChoice***.** A arquitetura do sistema é composta por cinco módulos interdependentes: Carga da Base, Processamento, Polaridade, Classificação e Sumarização.

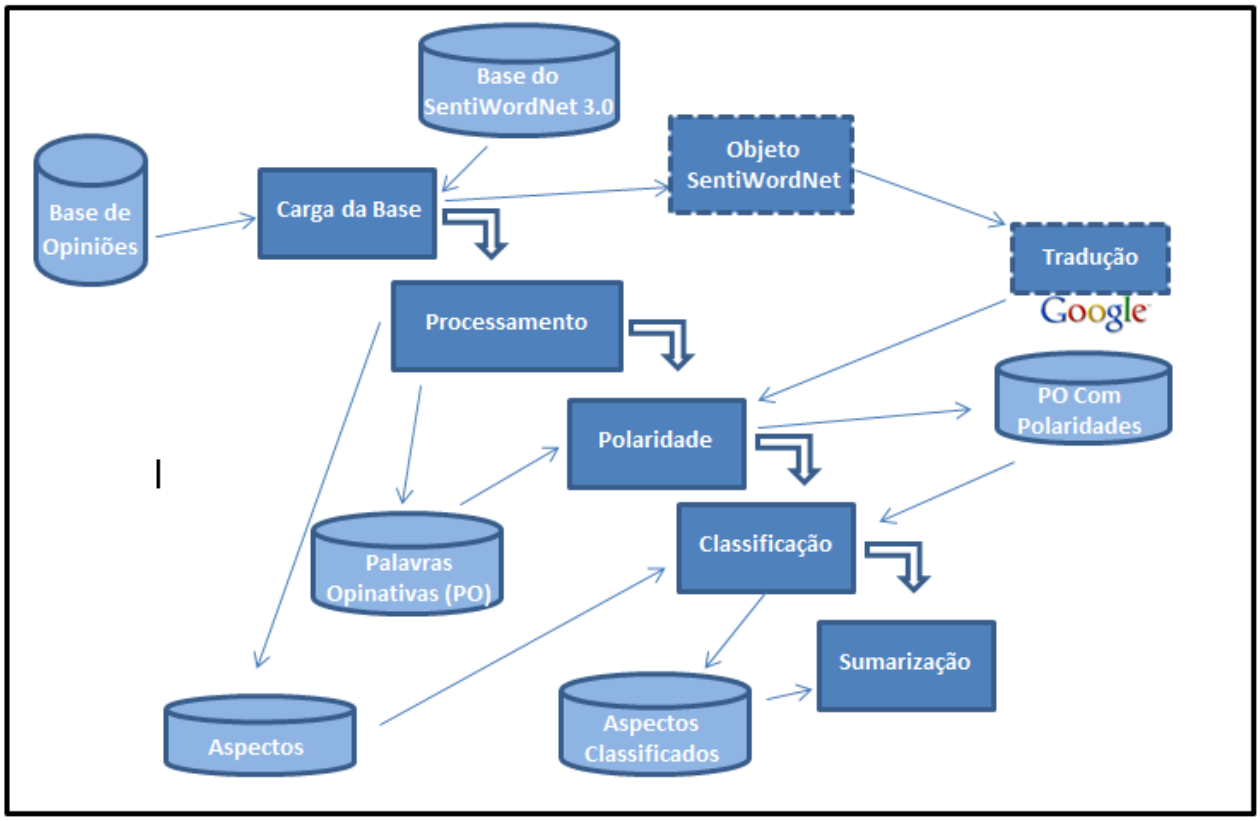

Figura 4.1: Arquitetura do BestChoice

Nas próximas seções, serão mostrados, em detalhes, cada um dos módulos, descrevendo as técnicas, os algoritmos e bases utilizadas para construção do sistema.

As figuras 4.2 e 4.3 mostram as telas principais do BestChoice, que oferecem os dois tipos de consultas possíveis do sistema:

- consulta por completo, onde, o sistema faz toda a análise sem a intervenção do usuário; e
- consulta parcial, onde o usuário poderá acompanhar e interferir no resultado de alguns dos módulos.

O segundo modo de consulta será a abordagem utilizada para explicação do sistema, que foi também a abordagem usada em [FERNANDES, 2010], e foi a abordagem usada para os testes.

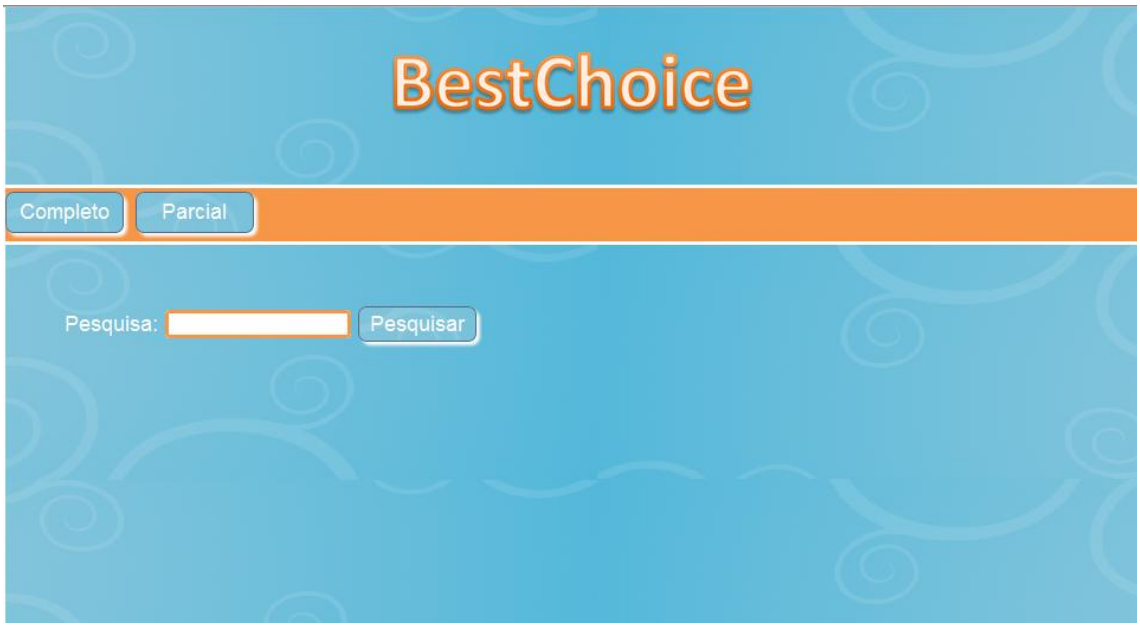

Figura 4.2: Tela Principal. Consulta por Completo

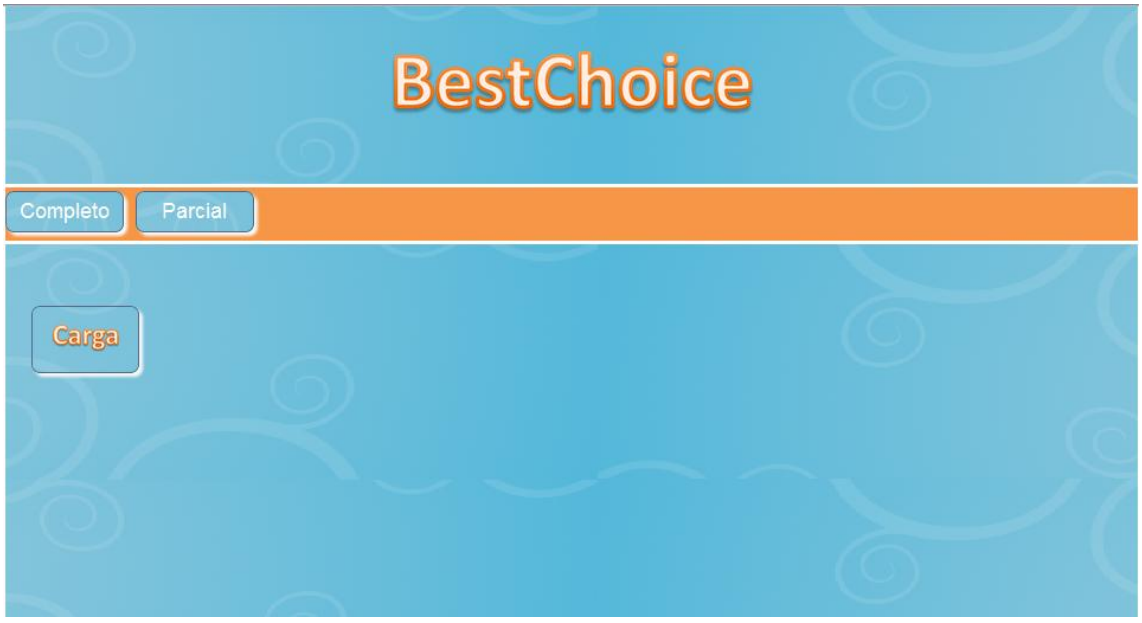

Figura 4.3: Tela Principal. Consulta Parcial

4.3. Bases Externas

O sistema se utiliza de duas bases externas: Base de opiniões e a Base do SentiWordNet.

A Base de Opiniões utilizada foi obtida a partir de textos do Twitter com escritas de até 140 caracteres. Foi utilizada a mesma base de dados obtida por Fernandes [FERNANDES, 2010], para que seja possível comparar os resultados obtidos pelos dois sistemas. A base e os resultados do trabalho [FERNANDES, 2010] foram fornecidos pelo próprio autor. Essa base é um arquivo de texto, onde cada linha (registro) contém uma palavra que representa o nome do bar/restaurante escolhido, seguida por uma tabulação, e o comentário a respeito do estabelecimento sob análise.

"Nome do Restaurante"	*"comentário"*

Por exemplo:

Spettus	*A comida do Spettus pode ser boa, mas o atendimento deixou a desejar.*

Exemplo 4.1: Exemplo criado pelo autor.

Apesar das opiniões estarem em um arquivo de texto, foi também implementada aqui a possibilidade de obter os textos diretamente da rede social do Twitter, através de uma API [TWITTER, 2010], que possibilita ao desenvolvedor consultar uma quantidade limitada de tweets. Porém, a API do Twitter foi unicamente utilizada para verificar o seu funcionamento no sistema, todos os testes realizados foram baseados nos comentários do Twitter fornecidos por Fernandes.

A segunda base externa utilizada foi a do SentiWordNet (seção 3.3). Foram utilizadas aqui mais de 110 mil palavras da base, cada uma com os seus scores positivos e negativos (ver Capítulo 5).

4.4. Módulo de Carga das Bases

Esse é o primeiro e mais simples módulo do sistema. Ele é responsável por carregar no sistema o arquivo com as opiniões, o arquivo com a base do SentiWordNet, o arquivo com as cláusulas de negação e o arquivo com as cláusulas adversativas.

As cláusulas adversativas e de negação serão utilizadas pelos módulos de classificação e polaridade. A partir da base do SentiWordNet será criado um objeto, que também será usado no módulo de Polaridade, para determinar a orientação, como positiva ou negativa, de cada uma das palavras opinativas.

A figura 4.4 representa a tela principal. Ela contém o botão [Carga], que dá início ao módulo. Após iniciar o módulo, teremos como resultado a figura 4.5.

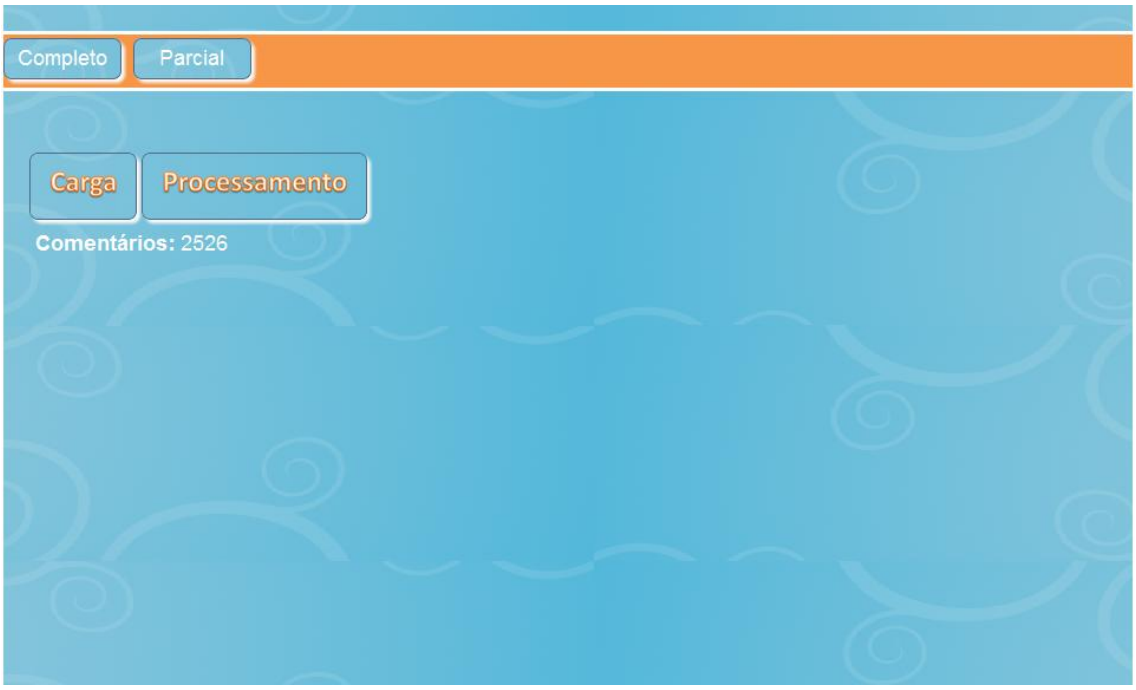

Figura 4.4: Resultado da Carga

A figura 4.4 mostra o resultado da carga dos dados. Como pode ser visto na imagem, a quantidade total de comentários (ou opiniões) analisados foram 2526.

4.5. Processamento

Logo após a carga das opiniões a serem analisadas, os textos são pré-processados, para que seja possível a execução do Pos-Tagger, bem como o levantamento das palavras opinativas e dos aspectos encontrados no texto.

Nas duas seções seguintes serão mostrados os processos para a aplicação do Pos-Tagger e para o levantamento dos termos do domínio.

4.5.1. POS-Tagger

Pos-Taggers, ou etiquetadores, são ferramentas que possuem como objetivo a identificação da classe gramatical correta de um vocábulo de acordo com o seu contexto. Neste trabalho, foi utilizado o mesmo Pos-Tagger usado em [FERNANDES, 2010], o Tree-Tagger [GAMALLO, 2005].

O Tree-Tagger é um etiquetador que pode ser usado para diversas línguas, inclusive para a língua portuguesa. Essa biblioteca recebe como entrada as sentenças, e retorna um conjunto de metadados com o lema e a classe gramatical de cada uma das palavras. Para exemplificar o uso do Tree-Tagger, considere a seguinte sentença:

A comida do Spettus pode ser boa, mas o atendimento deixou a desejar.

A saída produzida pelo Tree-Tagger será a seguinte:

Palavra	Classe	Lema
a	DET	a
comida	NOM	comida
do	PRP+DET	de
spettus	NOM	<unknown>
pode	V	poder
ser	V	ser
boa	ADJ	bom
,	VIRG	,
mas	CONJ	mas
o	DET	o
atendimento	NOM	atendimento
deixou	V	deixar
a	PRP	a
desejar	V	desejar
.	SENT	.

Tabela 4.1: Tabela com o resultado do processamento do Tree-Tagger

Porém, antes de executar o Tree-Tagger, todos os textos são agrupados e colocados em um formato aceito pela biblioteca. Esse formato contém uma palavra por linha, ou seja, cada linha deve conter uma palavra, ou pontuação, a ser etiquetada pelo Tree-Tagger.

No próximo passo, será utilizado o resultado obtido pelo Tree-Tagger, para o levantamento dos Termos.

4.5.2. Termos

Os termos são as palavras que serão classificadas (os aspectos) e as palavras que irão atribuir polaridade aos aspectos (as palavras opinativas).

Logo após a execução do Tree-Tagger, temos como resultado a tabela 4.1 com as palavras e classe gramatical associada a cada uma delas. Com a informação das classes gramaticais das palavras, são separados os aspectos, correspondentes aos substantivos (NOM), e as palavras opinativas, correspondente aos adjetivos (ADJ) e advérbios (ADV).

Durante o processo de levantamento dos termos, dois valores são calculados: o espectro de influência de cada um dos aspectos; e a frequência de cada uma das palavras.

- **Frequência**

A frequência será importante para a determinação da relevância dos termos. Os termos em cada classe (os adjetivos, advérbios e os substantivos) serão ordenados segundo sua frequência. A cada novo documento adicionado à base, uma nova execução deste módulo será necessária, para que esses valores de frequência sejam atualizados. A frequência será usada para determinar a ordem na qual os termos serão exibidos para o usuário (em ordem decrescente da frequência), e para aplicação do PMI.

- **Espectro de Influência**

O próximo passo refere-se ao cálculo do espectro de influência de cada um dos aspectos. Para o cálculo desse valor, adotamos o modelo seguido por [FERNANDES, 2010]. Esse valor irá corresponder à distância com que as palavras opinativas irão influenciar cada aspecto. Por exemplo, considere a seguinte frase:

A carne estava muito saborosa, porém a sobremesa estava terrível.

Exemplo 4.2: comentário criado pelo autor.

Na frase acima, considerando os aspectos *carne* e *sobremesa*, caso o espectro de influência tenha o valor três, o aspecto *saborosa* irá ser influenciado apenas pelo termo *saborosa*, e não pelo termo *terrível*, já que este possui uma distância maior que três para o aspecto *carne*.

Para se calcular o espectro de influência, inicialmente é montada uma sub lista para cada aspecto das palavras opinativas que aparecem ao redor deles. A fórmula para calcular a distância média para um aspecto candidato a_i é dada pela Equação 1, com *i* variando de *1* a *t*, onde *t* é o número total de aspectos na base.

$$dm(a_i) = \sum_{j=1}^{n} d(a_i, po_j) / n \qquad \text{(Equação 4.2)}$$

Onde po_j (com *j* variando de 1 a n) representa cada palavra opinativa na mesma sentença que o aspecto a_i, e *n* é o número de palavras opinativas na mesma sentença do aspecto a_i.

O próximo passo é calcular a média geral das distâncias (Equação 4.2), calculando o somatório das médias individuais de cada um dos aspectos, e dividindo por *t*, que corresponde ao número total de aspectos na base.

$$dm - global = \sum^{m} dm(a_i) / t \qquad \text{(Equação 4.2)}$$

Após calcular a frequência e o espectro de influência, a tela na Figura 4.5 é exibida para o usuário mostrando o total de substantivos, adjetivos e advérbios diferentes encontrados em todo o corpus do texto. Esse dados são unicamente informativos para o usuário.

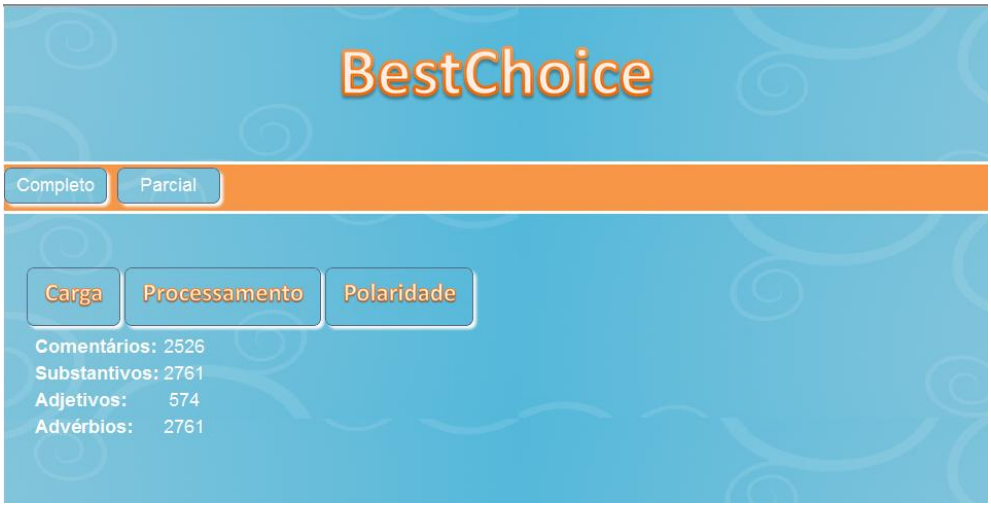

Figura 4.5: Resultado do Módulo Processamento

No próximo passo determinamos a polaridade de cada uma das palavras opinativas e quais termos serão usados para análise de sentimentos.

4.6. Polaridade

Nesse módulo será atribuída a polaridade a cada uma das palavras opinativas encontradas no texto. Esse é principal ponto no qual este trabalho se diferencia do método proposto por [FERNANDES, 2010]. Ao invés de darmos ao usuário a tarefa de atribuir a polaridade de cada uma das palavras opinativas, essa atribuição é feita pelo sistema e repassada para o usuário, que poderá modificar a polaridade das palavras se achar necessário.

Como dito, a atribuição da polaridade é realizada com o uso do SentiWordNet (SWN). Como a base do SentiWordNet é em inglês, utilizamos o tradutor do Google [GOOGLE-TRANSLATE, 2010], para realizar a consulta na base. Cada palavra opinativa é traduzida para o inglês e é feita uma consulta na base do SentiWordNet dos três valores referentes à palavra: o score neutro (Obj), o score positivo (Pos) e o score negativo (Neg). É atribuída à palavra a polaridade com maior score. Caso ocorra empate entre os dois scores maiores, será atribuída a polaridade Neutra para a palavra.

Esse módulo oferece a possibilidade do usuário modificar as polaridades atribuídas pelo sistema. Através da tela na figura 4.6, o usuário pode modificar ou confirmar as polaridades das palavras opinativas.

Figura 4.6: Tela para Alterar Polaridades

A seguir, será exibida a tela na figura 4.7, para que o usuário possa selecionar os termos correspondentes aos aspectos.

Figura 4.7: Tela de Seleção de Aspectos

Na próxima seção, será mostrado o módulo de Classificação, onde são atribuídas as polaridades aos aspectos.

4.7. Classificação

Após a identificação das palavras opinativas, de suas polaridades, e dos aspectos a considerar, esse módulo do sistema será responsável por determinar a polaridade dos aspectos selecionados pelo usuário no módulo anterior.

A tarefa de classificação das opiniões é dividida em quatro passos:

- identificação das palavras opinativas em cada sentença e suas polaridades;
- tratamento das expressões negativas;
- identificação e tratamento das expressões adversativas; e
- ligação entre a palavra opinativa e o aspecto.

As atribuições das polaridades de palavras opinativas já foram feitas no módulo anterior. No primeiro passo desse módulo serão buscadas no documento de entrada as palavras opinativas em cada sentença. Durante esse primeiro passo, uma sentença será classificada como no exemplo 4.3 a seguir:

*A carne não estava ruim [**-1**], mas dizer que a sobremesa estava maravilhosa [**+1**] é demais.*

Exemplo 4.3: comentário criado pelo autor.

Na sentença acima os adjetivos *ruim* e *maravilhosa* são classificados com suas polaridades iniciais. O advérbio *demais* também deveria ser considerado, porém para simplificar o exemplo ele não será contado.

No próximo passo, serão identificadas as cláusulas negativas e realizado o devido tratamento dessas expressões. As expressões negativas invertem o sentimento de um adjetivo ou advérbio. O processo para tratamento dessas expressões segue o método proposto por [FERNANDES, 2010], que considera que a polaridade da primeira palavra opinativa após a cláusula de negação deve ser invertida. Considerando o exemplo 4.3, a primeira polaridade será invertida por ser a primeira palavra opinativa após a cláusula de negação *não.* A polaridade do exemplo 4.3 ficará da seguinte forma (exemplo 4.4):

*A carne não estava ruim [**+1**], mas dizer que a sobremesa estava maravilhosa [**+1**] é demais.*

Exemplo 4.4: exemplo 4.3 com tratamento de expressões negativas.

As cláusulas de negação consideradas são mostradas na tabela 4.2, e são as mesmas cláusulas consideradas por em [FERNANDES, 2010].

Advérbios / Locuções adverbiais de Negação	não, nunca, jamais, nada, zero, tampouco, nem, nem um pouco, nem sequer, sequer, de modo algum, de jeito nenhum, de forma nenhuma, de forma Alguma

Tabela 4.2: Tabela com Advérbios e Locuções Adverbiais de Negação

No terceiro passo são tratadas as expressões adversativas. Essas expressões podem interferir bastante no sentimento atribuído a uma opinião, principalmente quando se trata de textos com conteúdos de ironias.

Segundo [ROCHA, 1994], as adversativas "relacionam pensamentos contrastantes". Para [GARCIA, 1992], as adversativas marcam oposição "às vezes com um matiz semântico de restrição ou ressalva". Já [CEGALLA, 1990] apresenta diversos sentidos para as adversativas, que para o autor elas "exprimem oposição, contraste, ressalva, compensação". De fato, as cláusulas adversativas exprimem um sentimento oposto entre duas sentenças.

Para tratar as cláusulas adversativas, verificamos a polaridade da sentença anterior à cláusula, e a polaridade da sentença após a cláusula. Caso sejam opostas, a polaridade atribuída à opinião que se encontra após a cláusula é invertida. Por exemplo, a polaridade vista no exemplo 4.4 ficará da seguinte forma (exemplo 4.5):

*A carne não estava ruim [**+1**], mas dizer que a sobremesa estava maravilhosa [**-1**] é demais.*

Exemplo 4.5: exemplo 4.4 após o tratamento de cláusulas adversativas

No exemplo acima, os adjetivos *ruim* e *maravilhosa* tinham a mesma polaridade após o passo anterior. Porém, como se encontrava uma conjunção adversativa "*mas"* entre as duas sentenças que continham os adjetivos, então o segundo adjetivo tinha sua polaridade invertida.

As expressões adversativas consideradas encontram-se na tabela 4.3.

Cláusulas Adversativas	mas, porém, todavia apesar disso, no entanto entretanto.

Tabela 4.3: Cláusulas Adversativas

O ultimo passo da classificação é a ligação entre os aspectos e as palavras opinativas. Para esse passo será usado o espectro de influência mostrado anteriormente. Para cada palavra opinativa, determina-se o aspecto ligado a ela - esse aspecto será aquele que se encontra a uma distancia menor ou igual ao espectro de influência médio. Logo, considerando o espectro de influência igual a 3, cada aspecto terá a polaridade mostrada no exemplo 4.6:

*A carne **[+1]** não estava ruim, mas dizer que a sobremesa [**-1**] estava maravilhosa é demais.*

Exemplo 4.6: exemplo 4.5 após a verificação do espectro de influência das palavras opinativas nos aspectos.

O próximo módulo é responsável por exibir os resultados da Análise de Sentimentos para o usuário. Na próxima seção esse módulo será mostrado em detalhes.

4.8. Sumarização

A sumarização é responsável por exibir ao usuário de forma simples e clara os resultados da classificação das opiniões. No caso do BestChoice, dois resultados são obtidos: o primeiro exibe ao usuário o resultado da classificação para que ele possa analisar os resultados de forma simples e clara; o segundo não é exibido para o usuário, pois serve apenas para uma análise comparativa com os resultados obtidos em [FERNANDES, 2010].

4.8.1. Resultado Usuário

No resultado para o usuário, são exibidos os aspectos e a porcentagem positiva e negativa para cada um dos aspectos. A figura 4.8 mostra a tela com o resultado da análise feita do restaurante Spettus.

Figura 4.8: Resultado mostrado para o usuário.

A figura 4.8 mostra o resultado análise para o restaurante Spettus. Três colunas são mostradas: a primeira mostra o aspecto analisado; a segunda e a terceira coluna mostram, respectivamente, a porcentagem de análises positivas e negativas, para cada aspecto. Para descobrir a porcentagem de análises neutras, basta subtrair de 100% a soma da análise positiva e negativa.

Neste capítulo foi mostrado cada passo para o desenvolvimento do sistema proposto, e os resultados conseguidos com a execução do sistema. No próximo capítulo, será mostrada a metodologia e os resultados dos experimentos.

5. Experimentos e Resultados

Nesse capítulo serão mostrados os resultados obtidos, a metodologia utilizada para realização dos experimentos e uma breve análise dos resultados.

5.1. Metodologia Para Experimentos

Como foi dito anteriormente, a base de dados usada para os experimentos foi disponibilizada por Fernandes, para que assim possa realizar comparações com os resultados obtidos em [FERNANDES, 2010]. No total foram disponibilizados 2526 comentários, dos quais tinham 51 marcas de bares e restaurantes.

Um segundo documento foi disponibilizado por Fernandes para que seja possível realizar as comparações com os resultados obtidos do BestChoice. O documento consiste de todos os aspectos analisados no HowGood, com suas respectivas polaridades. Cada linha desse documento é expressa da seguinte forma (exemplo 5.1):

15796722584	ADJ	saboroso	saboroso	+1

Exemplo 5.1

O primeiro valor identifica o comentário no Twitter do qual foi extraído a palavra. O segundo valor identifica a classe, no exemplo acima a classe é o adjetivo (ADJ). Os outros dois valores representam a palavra extraída e o lema, respectivamente. E o ultimo valor identifica a polaridade atribuída.

Das classes encontradas no documento, apenas três foram consideradas: adjetivos (ADJ) e advérbios (ADV), por representarem as palavras opinativas; e os substantivos (NOM), por representarem os aspectos. Como a escolha dos aspectos nos dois sistemas é feito pelo usuário, pode acontecer dos aspectos contidos no resultado de um sistema não estar presente no resultado do outro sistema. Por esse motivo, durante os testes só foram considerados os aspectos encontrados no resultado dos dois sistemas.

Os experimentos foram todos comparativos, dos quais, na melhor hipótese, seria esperado um resultado igual ao encontrado no HowGood, já que a polaridade das palavras opinativas nesse sistema foi realizada manualmente, e no nosso foi automatizado.

Durante a etiquetagem do Tree-Tagger nos 2526 comentários, foram extraídos 2761 substantivos, 574 adjetivos e 63 advérbios. Após essa etiquetagem, os aspectos passam pela tela de seleção de aspectos (figura 4.7). E as palavras opinativas, após serem classificadas a partir do SentiWordNet, passam pela tela de alteração das palavras opinativas, figura 4.6. Porém, para verificar o resultado da atribuição da polaridade pela base do SWN, durante a tela de alteração das palavras opinativas, foram deixadas com as polaridades atribuídas pela SWN.

Na próxima seção serão mostrados os resultados e a análise feita a partir deles. Serão mostrados também, alguns resultados anteriores, que resultaram na melhora do sistema.

5.2. Resultados e Avaliação

Como o principal objetivo do trabalho está na melhora da performance, automatizando algumas tarefas tirando do usuário tarefas não triviais, como a classificação de um grande número de palavras opinativas, os resultados desse trabalho foram obtidos a partir de experimentos comparativos. Essas comparações são feitas em relação aos resultados obtidos em [MILLIANO, 2010], e fornecidos pelo próprio autor.

A tabela 5.1 mostra algumas palavras opinativas (adjetivos ou advérbios) obtidas a partir dos primeiros experimentos.

Palavra	BestChoice	HowGood
ferreira	0	1
fritas	0	1
meninas	0	1
santo	0	-1
natal	0	-1

Tabela 5.1: Tabela parcial dos primeiros resultados de palavras opinativas

Esse primeiro resultado mostra uma classificação errônea do Tree-Tagger de algumas palavras como adjetivo ou advérbio. Apesar do ocorrido, o SentiWordNet no BestChoice classificou esses resultados como neutro, portanto, a classificação errada do Tree-Tagger no módulo de processamento não ocasionou perda da precisão na classificação dos aspectos. Do contrário, no caso do HowGood, que se utilizou do mesmo etiquetado, o Tree-Tagger, porém a classificação era manual, ocorreu nesses casos uma atribuição errada das polaridades, o que pode ser comum em uma classificação manual, quando se trata de uma grande quantidade de palavras.

Verificando a classificação errada do Tree-Tagger para alguns casos, foi utilizado a API *RiTa.WordNet* [RITA.WORDNET, 2010], uma API que possibilita acesso à antologia do *WordNet*. Essa com essa antologia é possível verificar, os antônimos e sinônimos das palavras, entre outras coisas, inclusive a classe gramatical a qual pertencem. Com isso fica possível verificar se as palavras que são classificadas pelo etiquetador, são realmente adjetivos ou advérbios, melhorando ainda mais a classificação das classes gramaticais, e consequentemente os resultados finais.

Após essa melhora, o sistema obteve uma significativa melhora na classificação dos adjetivos, e na atribuição final das palavras opinativas pelo SWN. A tabela 5.2 mostra os resultados obtidos após essa melhora no sistema.

Palavra	BestChoice	HowGood
amigo	1	-1
animadinho	0	-1
brava	-1	-1
calorento	-1	-1
cansada	-1	-1
central	1	1
certo	1	1
cheio	1	-1
corrido	-1	-1
cremoso	1	1
decadente	-1	-1
decepcionada	-1	-1
demorado	-1	-1
direto	0	1
divertido	1	1
especial	1	1
final	-1	-1
grande	1	1
impressionante	1	1
insossa	-1	-1
intenso	1	1
lento	-1	-1
local	0	-1
mais	1	-1
malditas	-1	-1
movimentada	1	1
nada	-1	-1
novo	1	1
nova	1	1
paga	1	0
pesado	-1	-1
pouca	-1	-1
predileto	1	1
recheado	0	0
rico	1	1
simples	-1	-1
tarde	-1	-1
tradicional	0	1
únicas	1	1
variada	1	1

Tabela 5.2: Tabela com a as polaridades das palavras opinativas

A partir da tabela acima é possível verificar uma semelhança de 80% entre as palavras opinativas. Porém, alguns casos que diferem os resultados nos dois sistemas, devem ser reconsiderados.

Para se aplicar o SWN, é necessária a tradução para o inglês antes de usar a base. Mas, algumas palavras foram traduzidas erroneamente para o inglês ao se usar a api do Google [GOOGLE-TRANSLATE, 2010]. Por exemplo, a palavra *animadinho* foi traduzida para *fired*, o que ocasionou em uma atribuição errada da polaridade pelo o SentiWordNet. Uma solução para esse caso seria uma base, igual ao do SWN, para o português.

Com outras palavras ocorreu uma provável atribuição errada da polaridade no sistema HowGood. Por exemplo, a palavra *amigo* que passa um sentimento positivo, mas no HowGood foi classificada como negativo. Um caso que ocorreu provável por erro durante a classificação manual.

E por fim, um dos problemas encontrado para a classificação do SentiWordNet é por ele não tratar de domínios específicos. Como no caso da palavra *local* que para o domínio de Restaurantes pode ser classificada como positivo, porém o SWN classificou como neutro. Como é possível treinar a base do SentiWordNet, uma ótima solução seria treinar a base para domínios específicos.

A tabela 5.3 mostra os resultados dos aspectos para o objeto *Spettus*. Os resultados obtidos se mostram bastante próximos aos resultados obtidos no HowGood.

Marca: Spettus	HowGood		BestChoice	
Aspecto	Positivo	Negativo	Positivo	Negativo
Spettus	14,29%	2,13%	28,29%	0,00%
carne	22,00%	6,00%	29,80%	7,30%
gerente	16,67%	0,00%	17,00%	0,00%
massa	21,62%	2,70%	23,30%	3,80%
rodízio	0,00%	11,11%	0,00%	11,11%
comida	17,70%	0,00%	18,90%	0,00%
churrascaria	17,65%	2,94%	18,45%	3,12%
lagosta	11,11%	0,00%	12,13%	0,00%
lugar	27,27%	3,41%	28,30%	3,80%
entrada	33,33%	0,00%	35,00%	0,00%
atendimento	42,50%	5,00%	43,40%	5,00%
cerveja	34,78%	0,00%	37,50%	0,00%
jantar	25,50%	0,67%	23,00%	7,00%
sushi	25,81%	6,45%	32,00%	3,00%
Chopp	12,50%	8,33%	6,25%	9,00%
variedade	0,00%	60,00%	0,00%	60,00%
prato	18,18%	6,82%	12,18%	6,30%
picanha	31,25%	0,00%	29,50%	0,00%
profissional	40,00%	0,00%	40,00%	0,00%
cozinha	43,75%	0,00%	43,75%	0,00%
restaurante	18,18%	0,00%	15,18%	2,50%

Tabela 5.3: Resultado comparativo entre os aspectos do HowGood e BestChoice

Apesar dos aspectos mostrarem os resultados bastante próximos, alguns resultados se mostram muito diferentes. Esses casos podem ser explicados por dois fatores: o resultado da polaridade das palavras opinativas, que foi mostrado na tabela 5.2; e a diferença entre os comentários analisados, enquanto foi fornecido para o BestChoice 2526 comentários, e no HowGood foram analisados 7713 comentários.

6. Conclusão

Este trabalho apresentou o sistema BestChoice, um sistema para classificação de sentimentos em nível de aspectos, baseado no sistema HowGood proposto em [FERNANDES, 2010].

As próximas seções estão dividas da seguinte forma: na seção 6.1 são mostradas as principais contribuições do BestChoice; na seção 6.2 serão mostradas as dificuldades encontradas durante a pesquisa e o desenvolvimento do sistema; e na seção 6.3 serão mostradas possíveis melhoras no sistema e oportunidades futuras de pesquisas.

6.1. Principais Contribuições

Inicialmente este trabalho fez um levantamento teórico de Análise de Sentimentos, mostrando os principais conceitos e etapas para a classificação de sentimentos de forma clara e sintética. Foi mostrada cada uma das fases da AS, exemplificando com o que há de mais recente na literatura e os principais desafios e limitações encontradas.

Como maior contribuição deste trabalho, temos a automatização da principal tarefa de classificação dos aspectos: a atribuição da polaridade das palavras opinativas. Em diversos trabalhos, tomando como exemplo [FERNANDES, 2010], fazem essa tarefa de atribuição das polaridades de forma manual, o que é totalmente inviável, quando se trata de várias palavras a serem classificadas como positivas, negativas ou neutras.

Foi implementado o sistema seguindo a arquitetura e descrição apresentada por este trabalho, de forma a manter a principais características de qualidade de software, criando um sistema modularizado, o qual, futuramente, qualquer um dos módulos possa ser modificado sem interferir em outro módulo.

Outra contribuição está no fato do sistema ser escrito para a língua portuguesa, já que a maior parte das pesquisas em AS são específicos para a língua inglesa. E com simples modificações nos parâmetros do sistema, ele pode ser usado para outras línguas.

6.2. Dificuldades Encontradas

Uma das principais dificuldades encontrada foi a de encontrar ferramentas, bases de dados e ontologias para AS na língua portuguesa, sendo preciso em muitos casos usar um tradutor da língua portuguesa, o que ocasionava em alguns casos erro na tradução, e consequentemente erros durante o processo de classificação dos sentimentos. Por ser uma área de pesquisa bastante recente, ainda há pouca pesquisa nesta área de AS, principalmente para a língua portuguesa.

Outra dificuldade encontrada foi em relação ao curto tempo para a realização das pesquisas, elaboração e implementação do sistema, e realização dos experimentos. Por esse fato, não foi possível realizar experimentos mais específicos e melhoras no sistema. Porém, apesar do curto prazo, foi possível atingir todos os objetivos propostos inicialmente.

6.3. Trabalhos Futuros

Pela limitação de tempo e por não serem o foco principal da pesquisa, diversas melhorias ficaram em aberto para trabalhos futuros. Entre elas estão:

1. **Criar uma base do SentiWordNet para a língua portuguesa:** Uma das maiores dificuldades foi a necessidade de se usar um tradutor do inglês para o português, pois em alguns casos foi realizada uma tradução errada ocasionando erro na classificação. A criação de uma base do SWN para língua portuguesa iria melhorar consideravelmente os resultados obtidos.

2. **Automatizar o processo de extração de aspectos:** Com uma alteração no módulo *Processamento*, usando o modelo proposto em [SIQUEIRA, 2010], seria possível automatizar o passo para escolha dos aspectos e das palavras opinativas.

3. **Combinação de Pos-Taggers:** Uma combinação de Pos-Taggers para a língua portuguesa iria melhorar na classificação da classe gramatical das palavras, o que melhoraria a classificação entre palavras opinativas ou aspectos.

4. **A utilização de um corretor ortográfico:** A ocorrência de erros de português é comum, principalmente em fóruns e redes sociais. A ocorrência desses erros é um dos muitos problemas enfrentados pelo sistema e na AS em geral, já que normalmente é considerado que todos os textos estão escritos segundo a norma gramatical, o que não é totalmente correto.

5. **Classificação da subjetividade:** Para o trabalho proposto, foi considerado que todos os textos eram opinativos, logo não foi implementado um módulo para detecção da subjetividade dos comentários. Uma proposta seria implementar um módulo para a classificação da subjetividade, através do modelo sugerido em [HATZIVASSILOGLOU & WIEBE, 2000].

REFERÊNCIAS BIBLIOGRÁFICAS

[BROOKSHEAR, 2005] BROOKSHEAR, J.G. **Ciência da Computação: Uma Visão Abrangente.** Porto Alegre: Artmed Editora, 2005.

[CARVALHO et al., 2009] CARVALHO, P.; SARMENTO, L.;SILVA, M. J.; OLIVEIRA E. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" **;-). In TSA '09 Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.** New York - USA, p. 53-56. 2009.

[CEGALLA, 1994] CEGALLA, D. P. **Novíssima gramática da língua portuguesa.** São Paulo: Nacional, 37 ed., 1994.

[DING et al., 2008] DING, X.; LIU, B.; YU, P.S. A holistic lexicon-based approach to opinion mining. **In WSDM '08: Proceedings of the international conference on Web search and web data mining**. New York – USA, p. 231–240. 2008.

[ESULI & SEBASTIANI, 2006] ESULI, A.; SEBASTIANI, F. Sentiwordnet: a publicly available lexical resource for opinion mining**. In proceedings of the 5th conference on language resources and evaluation.** Genoa - Italy, p.417-422. Maio 2006.

[FARIA & ZUQUIM, 2005] FARIA, E.; ZUQUIM, V. Uma Análise Crítica Da Influência Da Linguagem Da Internet No Cotidiano Do Interlocutor**. In XXV Congresso da Sociedade Brasileira de Computação.** São Leopodo – Rio Grande do Sul p. 53-56. 2009.

[FELLBAUM, 1998] FELLBAUM, C. **WordNet: An Electronical Lexical Database.** Cambridge, MA: The MIT Press, 1998.

[FERNANDES, 2010] FERNANDES, F. **Um Framework para Análise de Sentimento em Comentários sobre Produtos em Redes Sociais.** Dissertação (Mestrado em Ciência da Computação) - Centro de Informática/UFPE. Recife. 2010.

[FRAZON & GONÇALVES, 2006] FRAZON, C.; GONÇALVES, M.. Blogs corporativos: nova ferramenta de comunicação empresarial e/ou uma realidade ainda pouco brasileira. **Rp em revista.** Salvador, ano 4, jul. 2006.

[GAMALLO, 2005] GAMALLO, P. **PoS Tree-Tagger for Portuguese and Galician. 2005**. Disponível em: < http://gramatica.usc.es/~gamallo/tagger.htm > Acesso em: 10 de dezembro de 2010.

[GARCIA, 1992] GARCIA, O. M. **Comunicação em prosa moderna.** Rio de Janeiro: FGV, 15.ed., 1992.

[GOOGLE-TRANSLATE, 2010] GOOGLE-TRANSLATE; **Google Translate API.** 2010. Disponível em: < http://code.google.com/intl/pt-BR/apis/language/translate/overview.html > Acesso em: 10 de dezembro de 2010.

[HATZIVASSILOGLOU & WIEBE, 2000] HATZIVASSILOGLOU, V.; WIEBE, J. M. Effects of adjective orientation and gradability on sentence subjectivity,

Proceedings of the 18th conference on Computational linguistics. Saarbrücken-Germany, p.299-305. Julho 2000.

[HU & LIU, 2004] HU, M.; LIU, B. Mining and summarizing customer reviews. **In KDD'04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.** New York – USA, p. 168–177. 2004.

[IPNEWS, 2010] IPNEWS; **As 10 tecnologias estratégicas para 2011. 2010.** Disponível em: <http://www.ipnews.com.br/voip/pesquisas/pesquisas/as-10-tecnologias-estrat-gicas-para-2011.html>. Acesso em: 10 dez. 2010.

[KAJI & KITSUREGAWA, 2007] KAJI, N.; KITSUREGAWA, M. Building lexico n for sentiment analysis from massive collection of HTML documents. **Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and ComputationalNatural Language Learning.** Prague, Czech Republic, p. 1075–1083, 2007.

[KOTSIANTIS et al., 2007] KOTSIANTIS, S.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: a review of classification techniques. **Artificial intelligence review.** New York - USA, ano 3, v. 26, p.159-190. Novembro 2006.

[LIU et al., 2005] LIU, B.; HU m.; CHENG J. Opinion observer: analyzing and comparing opinions on the web**. In WWW '05: Proceedings of the 14th international conference on World Wide Web.** New York – USA, p. 342–351. 2005. ACM.

[LIU et al., 2009] LIU, B.; NARAYNAN, R.; CHOUDHARY, A. Sentiment Analysis of Conditional Sentences. **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.** Singapore, v. 1, p. 180–189. 2009.

[LIU, 2006] LIU, B. **Web data mining: Exploring Hyperlinks, Contents and Usage Data.** Chicago - Usa: Springer, 532 p. 2006.

[LIU, 2010] LIU, B.; **Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing.** Flórida – USA, 2a ed, Chapman and Hall/CRC, 2010.

[LOPES et al., 2008] LOPES, T.J.P.; HIRATANI, G. K.; BARTH, F.J.; RODRIGUES, O; MARACCINI, J. Mineração de Opiniões aplicada à Análise de Investimento. **WebMedia'08 Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web.** New York – USA, p. 117-120. 2008.

[MAGALHÃES, 2008] MAGALHÃES, L. H. **Uma análise de ferramentas para mineração de conteúdo de páginas web.** Dissertação (Mestrado em Engenharia) - COPPE/UFRJ. Rio de Janeiro. 2008.

[MAGALHÃES, 2009] MAGALHÃES, T. M. **Uma metodologia de mineração de opiniões na web.** Tese (Doutorado em Engenharia) - COPPE/UFRJ. Rio de Janeiro. 2009.

[MAUÁ, 2009] MAUÁ, D. **Modelos de Tópicos na Classificação Automática de Resenhas de Usuários.** Dissertação (Mestrado em Engenharia) – USP. São Paulo. 2009.

[MIHALCEA et al., 2007] MIHALCEA, R.; BANEA, C.; WIEBE, J. Learning multilingual subjective language via cross-lingual projections. **In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.** Prague, Czech Republic, p. 976-983. 2007.

[NIELSEN, 2009] NIELSEN; Global Faces and Networked Places - **A Nielsen report on Social Networking's New Global Footprint.** New York - USA. Nielsen 2009.

[OUNIS et al., 2006] OUNIS, I.; RIJKE, M.; MACDONALD, C.; MISHNE, G.; SOBOROFF, I. Overview of the TREC-2006 blog track. **In Proceedings of the 15th Text Retrieval Conference (TREC).** NIST - USA, p.15-27. 2006.

[PANG & LEE, 2008] PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends in Information Retrieval**, v.2 n.1-2, p.1-135, Janeiro 2008.

[PANG et al., 2002] PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? sentiment classification using machine learning techniques**. In Proceedings of the 2002 conference on empirical methods in natural language processing.** New Jersey – USA, v. 10, p. 79-86. 2002.

[POPESCU & ETZIONI, 2005] POPESCU, A.M.; ETZIONI, O. Extracting product features and opinions from reviews. **In HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.** Morristown - New Jersey - USA, p. 339–346, 2005.

[RITA.WORDNET, 2010] RITA.WORDNET; **WordNet library for Java/Processing.** 2010. Disponível em:<http://www.rednoise.org/rita/wordnet/documentation/index.htm> Acesso em: 10 de dezembro de 2010.

[ROCHA, 1994] ROCHA LIMA, C. H. **Gramática normativa da língua portuguesa.** Rio de Janeiro: José Olympio, 32. ed., 1994

[RODRIGUES, 2009] RODRIGUES, D. H. **Construção Automática de um Dicionário Emocional para o Português.** Dissertação de Mestrado - UBI, Covilhã, Portugal. 2009.

[SANTORINI, 1991] SANTORINI, B. Part-of-Speech Tagging Guidelines for the Penn Treebank Project. **Technical report MS-CIS-90-47**, University of Pennsylvania, Department of Computer and Information Science, 1991.

[SANTORINI, 1995] SANTORINI, B. **Part-of-Speech Tagging Guidelines for the Penn Treebank Project.** Philadelphia - Pennsylvania – USA.1995.

[SEBASTIANI, 2002] SEBASTIANI, S. Machine learning in automated text categorization. **ACM Computing Surveys.** New York - USA, v. 34, Issue 1, p. 1-47, Março 2002.

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

[THOMAS & COVER, 1991] THOMAS, J. A.; COVER, T. M. Elements of information theory. **Wiley-Interscience,** New York – NY - USA, 1991.

[TURNEY, 2002] TURNEY, P.; Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. **ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.** New Jersey – USA, p. 417 – 424. 2002.

[TURNEY, 2002] TURNEY, P.; Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. **ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.** New Jersey – USA, p. 417 – 424. 2002.

[TWITTER, 2010] TWITTER; **Twitter Search API.** 2010. Disponível em: <http://dev.twitter.com/doc/get/search > Acesso em: 7 de outubro de 2010.

[WIEBE & MIHALCEA, 2006] WIEBE, J.; MIHALCEA, R. Word sense and subjectivity. **Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.** Sydney, Australia, p. 1065-1072, Julho 2006,.

[WIEBE & WILSON, 2002] WIEBE, J.; WILSON, T. Learning to disambiguate potentially subjective expressions. **Proceedings of the Conference on Natural Language Learning (CoNLL).** New Jersey – USA , v. 20, p. 1–7, 2002.

[WIEBE et al., 2004] WIEBE, J. M.; WILSON, T.; BELL, M.; MARTIN, M. **Learning subjective language. Computational Linguistics.** Cambridge - Massachusetts - USA, v. 30, Issue 3, p. 277–308, September 2004.

[WILSON et al., 2004] WILSON, T.; WIEBE J.; HWA, R. Just how mad are you? Finding strong and weak opinion clauses. **AAAI'04 Proceedings of the 19th national conference on Artifical intelligence.** San Jose – Califórnia – USA, p. 761–767, 2004.

[WITTMANN & RIBEIRO, 1998] WITTMANN, L. H.; RIBEIRO, R. D. Recursos Linguísticos e Processamento Morfológico do Português: o palavroso e o projecto LE-PAROLE. **In: Lima, V.L.S. (ed.) Anais do II Encontro para o Processamento Computacional do Português Escrito e Falado.** Porto Alegre: Todeschini, p. 109-117. 1998.