Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Abordagem TOP(X) para Inferir os Comentários mais Importantes sobre Produtos e Serviços

Rogério Figueredo de Sousa

Teresina-PI, Agosto de 2015

Rogério Figueredo de Sousa

Abordagem TOP(X) para Inferir os Comentários mais Importantes sobre Produtos e Serviços

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Raimundo Santos Moura

Teresina-PI

Agosto de 2015

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco

S725a Sousa, Rogério Figueredo.
Abordagem TOP (X) para inferir os comentários mais importantes sobre produtos e serviços / Rogério Figueredo de Sousa. – Teresina: 2015.
79 f.: il.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Piauí, Teresina-PI, 2015.
Orientação: Prof. Dr. Raimundo Santos Moura.

1. Processamento de Linguagem Natural. 2. Análise de Sentimentos. 3. Mineração de Opiniões. 4. Lógica Fuzzy. I. Titulo.

CDD 006.35

...

Rogério Figueredo de Sousa

Abordagem TOP(X) para Inferir os Comentários mais Importantes sobre Produtos e Serviços

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação.

Raimundo Santos Moura
Orientador

Thiago Alexandre Salgueiro Pardo
Membro Externo

Pedro de Alcântara dos Santos Neto
Membro Interno

Ricardo de Andrade Lira Rabêlo
Membro Interno

Teresina-PI
Agosto de 2015

Aos meus pais Rubem e Rosa,
por sempre estarem comigo em todos os momentos.

Agradecimentos

Agradeço antes de tudo a Deus pois sem Ele eu nada poderia fazer.

Agradeço aos meus pais, Rubem e Rosa, pelo apoio incondicional dado em todos os momentos da minha vida.

Agradeço a minha noiva Milena Carvalho, por todo o amor, carinho e atenção a mim concedidos durante todo esse tempo.

Agradeço ao meu orientador, Prof. Dr. Raimundo Moura, por todos os conselhos, pela grande paciência e ajuda em tantos anos de orientação.

Agradeço aos amigos de mestrado, pelo companheirismo, por todos os momentos de aprendizado e também pelos momentos de descontração.

À todos os professores do Departamento de Computação da Universidade Federal do Piauí que contribuíram para minha formação, ao Prof. Dr. Ricardo Lira, que teve papel importantíssimo para que eu alcançasse essa vitória.

À CAPES/FAPEPI pelo apoio financeiro para realização deste trabalho de pesquisa.

Enfim, a todos que contribuíram direta ou indiretamente para a conclusão deste trabalho.

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes"

(Marthin Luther King)

Resumo

No contexto do constante crescimento da Web, diversos serviços foram virtualizados, incluindo o surgimento do comércio eletrônico (*e-commerce*). Tanto tradicionalmente quanto através de *e-commerce*, as pessoas necessitam comparar produtos e serviços para nortear suas decisões por meio da análise das características desejadas. A mudança que a Web ocasionou foi a exposição de suas opiniões em sites de compra e venda, fóruns na Web, redes sociais ou ainda grupos de discussão, permitindo sua visualização por qualquer pessoa que necessite. Porém, com o grande crescimento da Web, a quantidade de dados também aumentou, impossibilitando a coleta manual dessas opiniões. Logo, a busca automática de opiniões, promoveu o crescimento da área de Análise de Sentimentos, a qual é responsável por criar técnicas automáticas para coletar, analisar e sumarizar as opiniões encontradas em diversos locais na Web. Este trabalho apresenta a abordagem TOP(X) para estimar a importância de comentários disponibilizados por clientes na Web, por meio do uso de um sistema *Fuzzy*. O sistema *Fuzzy* possui três variáveis de entrada: reputação do autor, número de tuplas $\langle característica, palavra\ opinativa \rangle$ e a porcentagem de palavras escritas corretamente; e uma variável de saída: grau de importância do comentário. Realizou-se também um experimento para comparar os resultados de um método de orientação semântica executado sobre todos os comentários e sobre uma seleção dos melhores comentários de um *corpus*. O experimento foi conduzido com 1620 comentários sobre *smartphones* (982 positivos, 594 negativos e 44 negativos) e nossa abordagem melhorou os resultados do método de orientação semântica em aproximadamente 10% na medida *F-measure* para comentários positivos e 20% para comentários negativos.

Palavras-chaves: Processamento de Linguagem Natural. Análise de Sentimentos. Mineração de Opiniões. Lógica Fuzzy.

Abstract

In the context of the constant growth of the Web, several services have been virtualized, including the emergence of the e-commerce. Both traditionally and through e-commerce, people need to compare products and services to guide their decisions using the analysis of the desired features. The web allowed a major exposition of one's views through sales websites, web forums, social networks, and discussion groups, allowing their visualization by anyone who needs it. However, with the explosive growth of the Web, the amount of data has also increased, precluding the manual collection of these opinions. Thus, the automatic search for reviews, promoted the growth of the Sentiment Analysis field, which is responsible for creating automated techniques to collect, analyze and summarize the opinions gathered in several places on the web. This work presents the TOP(X) approach to estimate the importance of reviews provided by customers on the Web, using a fuzzy system. The Fuzzy system has three input variables: author reputation, number of tuples $\langle feature, opinionative\ word \rangle$, and percentage of correctly spelled words; and one output variable: reviews importance. Also, an experiment was conducted in order to compare the results of a semantic orientation method performed over all the reviews and over a selection of the best reviews in a corpus. This experiment was conducted with 1620 reviews about smartphones (982 positives, 594 negatives and 44 neutral) and our approach improved the results of the semantic orientation method up to approximately 10% in f-measure for positive reviews and 20% in f-measure for negative reviews.

Keywords: Natural Language Processing. Sentiment Analysis. Opinion Mining. Fuzzy Logic.

Lista de Figuras

Lista de tabelas

Lista de abreviaturas e siglas

ANTLR	ANother Tool for Language Recognition
EBNF	Extended Backus Naur Form
EI	Extração da Informação
FCL	Fuzzy Control Language
FIS	Fuzzy Inference System
FLC	Fuzzy Logic Controller
FOM	Fuzzy Opinion Mining
GLN	Geração de Linguagem Natural
LSA	Latent Semantic Analysis
NLIDB	Natural Language Interfaces to Databases
NLTK	Natural Language ToolKit
PLN	Processamento de Linguagem Natural
PMI	Pointwise Multual Information
P&R	Perguntas e Respostas
RSO	Redes Sociais Online
RI	Recuperação da Informação
TSK	Takagi-Sugeno-Kang
UFPI	Universidade Federal do Piauí

Sumário

Introdução

Neste capítulo são apresentados o contexto do presente trabalho, as motivações da pesquisa, os objetivos pretendidos, contribuições científicas, bem como a organização do presente texto.

Contexto e Motivação

A busca de referências ou opiniões é um ato comum entre as pessoas quando se tem interesse por produtos ou serviços. Além do interesse dos consumidores, existem empresas que produzem produtos ou disponibilizam serviços que também estão interessadas nas opiniões de seus consumidores e, por isso, buscam meios de analisar essas informações para nortear suas ações de marketing e tomadas de decisão.

Segundo Liu (2010), o interesse de analisar opiniões sempre existiu, seja por parte das pessoas ou empresas, porém, o grande crescimento da popularidade da Web modificou drasticamente a forma em que as opiniões são emitidas e disponibilizadas. Antes deste marco, as pessoas tinham acesso a opiniões apenas de pessoas próximas: familiares, amigos ou conhecidos. Já as empresas criavam enquetes ou pesquisas de opinião para obter a satisfação geral de seus consumidores.

Com a popularização da Web, as pessoas e empresas passaram a ter novos lugares para deixar e coletar opiniões. Blogs, fóruns e grupos de discussões são os principais locais que recebem uma grande quantidade de conteúdos gerados por usuários. Adicionalmente, empresas de compra e venda pela Web passaram a disponibilizar áreas em seus sites para receber comentários de clientes. Mais de 40% das pessoas no mundo moderno dependem de opiniões, revisões e recomendações de outros consumidores para comprar produtos ou requisitar serviços (KHAN et al., 2009; AL-MAIMANI; SALIM; AL-NAAMANY, 2014; PANG; LEE, 2008).

Nos últimos anos, surgiram as chamadas Redes Sociais *Online* (RSO) que incrementaram ainda mais a oferta de locais disponíveis para geração de conteúdo opinativo entre usuários de produtos e serviços. O desafio dos pesquisadores é extrair informações dos grandes volumes de dados não estruturados disponíveis na Web.

Embora haja conteúdo opinativo para os consumidores e empresas, a enorme quantidade de informações torna impraticável a análise manual deste conteúdo. Portanto, surgiu a necessidade do desenvolvimento de métodos e ferramentas automáticas para coletar, analisar e sumarizar o conteúdo opinativo em geral.

Nesse contexto, esta Dissertação de Mestrado apresenta uma abordagem para

estimar a importância de comentários opinativos escritos por usuários na Web sobre produtos ou serviços. A abordagem proposta é baseada em técnicas de Processamento de Linguagem Natural (PLN) e sistemas *fuzzy* (ZADEH, 1975a) e considera três variáveis de entrada: reputação do autor, quantidade de características avaliadas pelos consumidores e o percentual de palavras escritas corretamente.

A evolução dos trabalhos sobre coleta e análise de opiniões fez surgir o campo de estudo em Análise de Sentimentos, também conhecido como Mineração de Opiniões (LIU, 2010). Atualmente, essa campo ganha força tanto na academia quanto na indústria de comunicação e marketing.

A Análise de Sentimentos é definida como qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual e pode ser estruturada genericamente em três etapas (LIU, 2012):

1. Identificar as opiniões expressas sobre determinado assunto ou alvo em um conjunto de documentos;
2. Classificar a orientação ou polaridade dessa opinião, isto é, se tende a positiva ou negativa;
3. Apresentar os resultados da classificação de forma agregada e sumarizada.

Embora, para cada etapa, existam muitas pesquisas realizadas, ainda há a necessidade de muitos avanços para melhorar os resultados existentes na academia e na indústria. Existem limitações que ainda não possuem soluções satisfatórias, tais como reconhecimento de entidades nomeadas, resolução de anáforas, tratamento de opiniões negativas, resolução de ambiguidades, dentre outros. Embora alguns destes procedimentos sejam também comuns à área de PLN, eles atingem diretamente a problemas encontrados no campo de Análise de Sentimentos, por se tratar de uma subárea de PLN.

Devido às limitações ainda existentes, a maioria dos trabalhos da literatura está voltada para melhorar os resultados existentes aplicando técnicas e abordagens nas etapas conhecidas. Como exemplo, trabalhos como Zhang e Liu (2011), Silva, Lima e Barros (2012) concentram seus esforços na identificação das opiniões expressas nos documentos opinativos, recuperando características dos produtos ou ainda as opiniões dos usuários sobre estas características.

Outros trabalhos como os de Pang e Lee (2008), Turney (2002), além de apresentar métodos para identificar opiniões, apresentam esforços na definição da orientação semântica das opiniões identificadas nos documentos. Além disso, discutem formas de apresentação de seus métodos de análise, de forma sumarizada e amigável para os usuários dos seus sistemas.

Além dos trabalhos supracitados, existem outros que buscam contribuir com o desenvolvimento de recursos úteis para a comunidade acadêmica como: i) léxicos de sentimentos (WIEBE; RILOFF, 2005; BACCIANELLA; ESULI; SEBASTIANI, 2010; SOUZA et al., 2011; SILVA; CARVALHO; SARMENTO, 2012); ii) *features* para uso em técnicas de aprendizado de máquina (PANG; LEE; VAITHYANATHAN, 2002; DAVE; LAWRENCE; PENNOCK, 2003); e iii) *corpus* anotados de referência (FREITAS et al., 2012). Tais recursos apresentam evoluções importantes para o desenvolvimento da Análise de Sentimentos.

Objetivos

O principal objetivo deste trabalho consiste em estimar a importância de comentários de usuários sobre produtos ou serviços, utilizando técnicas de PLN, Análise de Sentimentos e Sistemas *Fuzzy*. Dessa forma, desenvolveu-se uma abordagem para extrair características dos comentários de clientes, e classificar (*ranking*) de acordo com a importância estimada. Argumenta-se que os comentários mais importantes devem resumir ou reunir as melhores opiniões dos usuários, permitindo que novos clientes possam avaliar apenas um pequeno conjunto de comentários e tomar suas decisões. Adicionalmente, os métodos automáticos de identificação de orientação semântica podem se beneficiar avaliando apenas uma parte de um grande conjunto de comentários, gerando melhores resultados em precisão e desempenho.

O objetivo principal pode ser detalhando nos seguintes objetivos específicos:

- Discutir métricas (variáveis) que possam ser utilizadas para ponderar a importância de comentários de consumidores sobre produtos ou serviços;
- Propor um método de inferência da orientação semântica baseado em léxico de sentimentos para ser utilizado na avaliação da abordagem proposta;
- Inferir a importância de comentários de consumidores, por meio da modelagem de um sistema de inferência *fuzzy*;
- Realizar um estudo experimental da abordagem e discutir os resultados obtidos.

É importante destacar que este trabalho integra um projeto mais amplo, em desenvolvimento no DC-UFPI, que engloba desde a extração de características de um produto em sites de fabricantes, a análise de comentários de usuários Web em sites de empresas que comercializam o produto ou realizam comparação de preços, até a análise das reclamações feitas sobre o produto ou empresa no site do Reclame-aqui.

Contribuições

Este trabalho foi validado pela comunidade acadêmica através da publicação dos seguintes trabalhos:

- *An Approach to Select the Best User Reviews on the Web* (CICLing, 2015): Este trabalho apresenta a versão preliminar da abordagem TOP(X), com a modelagem inicial do sistema *fuzzy* utilizado para a inferência de importância dos comentários. As três variáveis atuais de entrada do sistema foram apresentadas, porém, implementou-se somente duas variáveis: quantidade de tuplas e riqueza do vocabulário. Como a modelagem do sistema já previa o uso da variável de reputação do autor, e, no entanto, ela não foi implementada, o seu valor foi fixado em 1. Esse trabalho apresentou o procedimento de ajuste do sistema *fuzzy* de acordo com um *corpus* de referência com 350 comentários e um procedimento experimental sobre um conjunto de 1620 comentários. Apesar da ausência da variável de reputação do autor, o experimento desse trabalho apresentou bons resultados.
- *A Fuzzy System-Based Approach to Estimate the Importance of Online Customer Reviews* (FUZZ-IEEE, 2015): Essa é a versão completa do trabalho contendo todos os métodos apresentados na versão enviada anteriormente ao CICLING, porém, nessa oportunidade, foram acrescidas novidades em relação ao trabalho preliminar. Todas as três variáveis foram implementadas: reputação do autor, quantidade de tuplas e riqueza do vocabulário. A modelagem do sistema *Fuzzy* foi refeita de forma a permitir a inferência da importância de acordo com os novos valores das variáveis, principalmente com relação à reputação do autor. Além disso, o método de polarização utilizado na experimentação foi modificado, permitindo a manipulação de modificadores adverbiais sobre os sentimentos avaliados, diferenciando, dessa forma, construções como: "bateria boa" e "bateria *muito* boa", por meio de uma base de advérbios criada para esse fim. Adicionalmente às novidades desse trabalho, foi apresentado também o procedimento de experimentação atualizado, que avaliou os mesmos 1620 comentários obtendo bons resultados.

Organização desta Dissertação

Além deste capítulo introdutório, esta dissertação é composta de quatro outros capítulos, das referências bibliográficas e dos apêndices, conforme detalhamento a seguir:

No Capítulo 1 – Fundamentação Teórica, são apresentados conceitos e definições referentes à área de Processamento de Linguagem Natural (PLN), tais como: etapas do processo de interpretação de linguagens naturais, aplicações em PLN, conceitos linguísticos, dentre outros. Em seguida, discute-se também definições do campo de estudos em Análise

de Sentimentos, incluindo os níveis de granularidade que uma opinião pode ser analisada e os tipos de classificação de uma opinião. Em seguida, apresentam-se os conceitos e definições sobre Sistemas de Inferência *Fuzzy*. No final, discutem-se as ferramentas e recursos que foram utilizados neste trabalho, a saber: etiquetadores (*POS Tagger*), *parsers*, *lemas*, padrões linguísticos, *corpus* de comentários e o *corpus* de referência, que foi usado para avaliar a abordagem proposta.

No Capítulo 2 – Trabalhos Relacionados, apresenta-se uma revisão da literatura com os principais trabalhos das áreas de extração de opiniões e definição de orientação semântica. No final, discutem-se também os trabalhos que envolvem sistemas *fuzzy* na área de mineração de opiniões.

No Capítulo 3 – Abordagem TOP(X), descreve-se o modelo de um sistema de inferência *fuzzy* para definir a importância de um comentário feito por um usuário Web sobre um produto um serviço. O modelo proposto consiste de três variáveis de entrada: reputação do autor, quantidade de pares <característica, palavra opinativa> e a porcentagem de palavras escritas corretamente e uma variável de saída: grau de importância do comentário.

No Capítulo 4 – Experimentos e Resultados, são descritos os experimentos realizados para avaliar a abordagem proposta. Para cada experimento, apresenta-se o plano do experimento, os resultados obtidos e uma ampla discussão com a análise dos erros e ameaças externas ao experimento.

Finalmente, no último capítulo – Conclusões e Trabalhos Futuros, apresenta-se um balanço das contribuições apresentadas nesta dissertação, as dificuldades encontradas para a sua realização, bem como a indicação de alguns trabalhos que poderão ser desenvolvidos em um futuro próximo.

1 Fundamentação Teórica

Neste capítulo são apresentados os principais conceitos referentes a Processamento de Linguagem Natural, Análise de Sentimentos e Sistemas *Fuzzy*, além de ferramentas e recursos importantes para o entendimento geral deste trabalho.

1.1 Processamento de Linguagem Natural

De acordo com o dicionário da língua portuguesa (FERREIRA, 2004), o verbete linguagem é definido como "o uso da palavra articulada ou escrita como meio de expressão e de comunicação entre pessoas". Portanto, a linguagem constitui um dos aspectos fundamentais do comportamento humano, pois ela representa um veículo de transmissão e difusão de cultura entre os povos. Na forma escrita, ela serve para registrar conhecimentos que podem passar de geração para geração. Na forma falada, ele serve para a comunicação dia a dia entre as pessoas.

Entende-se por "Linguagem Natural" qualquer linguagem que é usada todos os dias para a comunicação das pessoas, tais como Inglês, Português ou Espanhol (BIRD; KLEIN; LOPER, 2009). O termo "Natural" é utilizado para distinguir a fala e escrita humana das linguagens mais formais, tais como notações matemáticas ou lógicas, ou linguagens de programação.

O conjunto de técnicas computacionais formais desenvolvidas para tratar dessas Linguagens Naturais é denominado Processamento de Linguagem Natural. Segundo Jackson e Moulinier (2007), o termo Processamento de Linguagem Natural (PLN) é normalmente utilizado para descrever a função dos componentes de software ou hardware em um sistema computacional que analisam ou sintetizam a linguagem escrita ou falada.

A área de PLN busca extrair uma completa representação do significado de textos livres. A grosso modo, equivale a descobrir em um texto os seis principais elementos usados para organizar os pensamentos: quem, o que, quando, como, onde e por quê. Esses elementos são oriundos do jornalismo investigativo e são referenciados como *6W (who, what, when, how, where e why)* (ROBERTSON, 1946). Para isso, a área de PLN frequentemente utiliza conceitos linguísticos como classes gramaticais (Substantivos, Verbos, Adjetivos) e estruturas gramaticais (Sintagmas Nominais, Sintagmas Verbais, Sintagmas Preposicionais), além de tratar de grandes desafios como resolução de anáforas e ambiguidades.

As técnicas de PLN fazem uso de várias representações de conhecimento, como léxicos de palavras e seus significados, propriedades gramaticais, conjuntos de regras gramaticais e frequentemente outros conjuntos de recursos como ontologias de entidades e

ações, ou ainda *Thesaurus* de sinônimos e abreviações (KAO; POTEET, 2007). Existem muitas aplicações para as técnicas de PLN, dentre elas algumas merecem ser comentadas: Tradução Automática do Chinês, Sumarização Automática de Textos, Recuperação e Extração da Informação (RI e EI, respectivamente), Perguntas e Respostas (P&R), Geração de Relatórios, Construção de Ontologias, Mineração de Textos Biomédicos, Categorização Textual e Análise de Sentimentos e Subjetividade (INDURKHYA; DAMERAU, 2010).

- A tradução automática do chinês é uma aplicação especial pois consiste de um conjunto de técnicas especialmente desenvolvidas para tratar da tradução automática da língua mais falada no mundo. Estima-se que existam 1.2 bilhões de pessoas que possuem o chinês como língua primária ou secundária, especificamente o *Mandarin*, enquanto o inglês, que está em segundo lugar neste *ranking*, possui cerca de 330 milhões de nativos e por volta de 150 milhões de pessoas que a utilizam como segunda língua. Além disso, o Mandarim padrão é adotado também como uma das seis línguas oficiais das Nações Unidas, portanto é classificada como uma das seis línguas mais influentes do mundo quando avaliada sob a ótica do total de falantes pelo mundo, a influência geográfica, o poder econômico dos países utilizando a língua e o uso literário e científico da linguagem. Consequentemente, dentre razões econômicas, políticas, culturais e humanitárias, a tradução automática do chinês (e para o chinês) é uma das mais importantes aplicações na área de PLN. Essa importância é justificada também pelo fato de que o desenvolvimento dos sistemas para tradução automática do chinês foi realizado paralelamente aos sistemas de tradução automática da área como um todo.

- A sumarização automática de textos, segundo (MANI, 2001), é o processo de extrair conteúdo de uma determinada fonte e apresentar as informações mais importantes aos usuários de forma condensada e sensível às necessidades dos interessados, sejam usuários ou aplicações. De acordo com o mesmo autor, a sumarização faz parte da vida diária das pessoas pois sempre há alguma forma de sumarização no dia a dia. Seja por meio da manchete de jornais, ao se narrar um evento a uma pessoa, resumos de textos científicos dentre outros. Devido a sua utilidade e frequência, automatizar o processo de sumarização tornou-se um grande campo de pesquisa na comunidade de PLN (MARTINS et al., 2001). É importante ressaltar que o campo de sumarização automática foi também afetada pelo crescimento da Internet, pois esse evento provocou um aumento considerável na quantidade de documentos disponíveis *online* e a na necessidade de se obter informações em larga escala. Em especial, a sumarização automática já possui diversas ferramentas para a língua portuguesa, por exemplo, GistSumm (PARDO; RINO; NUNES, 2003a), NeuralSumm (PARDO; RINO; NUNES, 2003b), SuPor (LEITE et al., 2007), entre outros. Outros sumarizadores automáticos podem ser encontrados no trabalho de Pardo (2008).

- Segundo Baeza-Yates, Ribeiro-Neto et al. (1999) Recuperação da Informação (RI) pode ser definida como a aplicação que lida com a representação, armazenamento, organização e o acesso a unidades de informação. Essas unidades de informação podem ser referências a documentos reais, os próprios documentos ou ainda parágrafos, páginas na Web, imagens, figuras, músicas, vídeos, etc. Porém, em se tratando de PLN, o foco da RI é principalmente em documentos escritos. Uma das mais importantes características da RI é o fato de que um sistema de Recuperação de Informações tem de lidar com dados imprecisos e incompletos, tanto em relação às necessidades dos usuários quanto ao conjunto no qual as informações são coletadas. Apesar disso, a RI pode ser considerada um sucesso de aplicação de PLN, principalmente, devido à popularização e livre oferta de ferramentas de busca na Internet.

- Diferentemente da Recuperação da Informação ao tratar de unidades informacionais como sendo quaisquer mídias, a Extração de Informação (EI) trata exclusivamente de descrições textuais. A EI é o processo de analisar textos em busca de informações úteis, por exemplo, entidades, relações ou eventos, em resumo: "quem fez o que com quem, quando e onde". A EI surgiu em resposta à necessidade de processamento eficiente de textos em domínios especializados, focando somente em partes relevantes de textos e ignorando o restante, ou seja, a EI busca capturar informações estruturadas sem sacrificar a viabilidade do processo.

- Outra aplicação das técnicas de PLN a ser destacada são os sistemas de Perguntas e Respostas (P&R), que podem ser definidos, em linhas gerais, como sendo um processo automático capaz de entender questões formuladas em uma linguagem natural como o português e assim responder exatamente o que foi solicitado. Um sistema de P&R "ideal" é bastante complexo, apesar da definição simples. Analisando como um fluxo de tarefas, um sistema de P&R deve ser capaz de determinar a informação expressa na questão, localizar a informação requerida, extraí-la, gerar uma resposta e apresentá-la de acordo com os requerimentos existentes na questão. Este processo em geral, requer a capacidade de interpretar questões e analisar documentos que estão escritos em linguagens naturas com domínio irrestrito, além de apresentar interação confortável e apropriada para os usuários. Diante disso, sistemas de P&R dependem do avanço de pesquisas em múltiplos campos, como RI e *Natural Language Interfaces to Databases* (NLIDB), portanto a comunidade acadêmica tem concentrado esforços na resolução de problemas nos diversos campos de estudo.

- A Geração de Relatórios é uma aplicação específica do domínio de Geração de Linguagem Natural (GLN). A Geração de Relatórios pode ser definida como a produção automática de material escrito para atender à solicitação de algum usuário. Para a geração deste material, normalmente são analisados dados estruturados (por exemplo, séries numéricas temporais), sendo que o tamanho dos relatórios podem ser

de um parágrafo ou várias páginas. Geradores de Relatórios foram desenvolvidos para muitos domínios, como, mercado de trabalho, mercado de ações, clima, qualidade do ar, dentre outros.

- Para tomar decisões o ser humano utiliza o conhecimento sobre o mundo que foi adquirido em diversos momentos de sua vida, por exemplo, leitura e conversação. O fato é que toda pessoa possui uma quantidade de conhecimento de como o mundo é estruturado e esse conhecimento é utilizado para guiar o dia a dia de cada um. Diante disso, o desafio a ser considerado é explicitar esse conhecimento de forma que possa ser compartilhado e utilizado para determinados fins. Dessa forma, a construção de ontologias tem o objetivo de capturar conhecimento sobre o mundo e explicitá-lo para uma proposta ou tarefa específica. Atualmente as ontologias são modestas e buscam limitar-se a capturar conhecimento sobre um determinado domínio, algum fenômeno ou até sobre determinada situação ou evento, porém, a proposta essencial das ontologias é capturar conhecimento sobre uma certa realidade de modo declarativo, independentemente da aplicação e o modo como esse conhecimento é usado, de tal forma que alguém possa reutilizar esse conhecimento e aplicá-lo em outro contexto.

- A aplicação de PLN em dados biomédicos é conhecida como BioNLP (processamento de linguagem biomédica, ou ainda, mineração de dados biomédicos). O ramo da biomedicina apresenta certos desafios pois possui um conjunto de dados e tarefas únicos em termos de base de dados para pesquisa, e possui diversos aspectos que são interessantes para a comunidade de PLN. Além de questões éticas que devem ser asseguradas, a qualidade dos softwares deve ter uma atenção além do normal em comparação às outras aplicações de PLN. Adicionalmente, a aplicação de técnicas no tratamento de textos biomédicos varia desde o uso de tokenização até o reconhecimento de entidades nomeadas para a construção de *corpus* e representação semântica.

- A Categorização Textual é o processo de categorizar documentos com o objetivo de organizá-los por temas de interesse para posteriormente encontrá-los de forma rápida e precisa. A categorização de textos não é uma disciplina recente, mas a sua modernização data do século XIX, com o crescimento das universidades. Segundo Barbosa (1969) a biblioteconomia é a área que mais se beneficiou da categorização textual. De acordo com Lima (2000) a categorização textual tornou-se um tema bastante difundido atualmente na área de sistemas de informação, devido ao crescimento da Web associado com o crescimento das bases de textos digitais nas empresas.

- Por fim, a Análise de Sentimentos ou Mineração de Opiniões é o estudo computacional de opiniões, sentimentos e emoções expressadas de forma textual. A principal tarefa

da Análise de Sentimentos é extrair informações úteis dos conteúdos gerados por clientes, e assim, definir o sentimento das suas opiniões. Porém, encontrar fontes de opiniões na web e monitorá-las é um grande desafio, por causa da imensa quantidade de fontes de textos opinativos. Portanto, é um problema desafiador da área de PLN e devido ao seu grande valor para aplicações práticas, houve um grande crescimento tanto na academia quanto na indústria.

Para dar suporte aos diversos tipos de aplicações de PLN, um sistema de linguagem natural deve usar considerável conhecimento acerca das estruturas da linguagem alvo, o que inclui saber o que são palavras, quais palavras devem ser combinadas para formar sentenças, o que as palavras significam, como o significado das palavras contribuem para o entendimento geral das sentenças, e assim por diante (ALLEN, 1995). Tradicionalmente, os trabalhos em PLN tendem a considerar o processo de análise da linguagem como uma decomposição em uma quantidade de estágios, espelhando as distinções da linguística teórica, que em resumo variam em sintaxe, semântica e pragmática (INDURKHYA; DAMERAU, 2010). Estes estágios podem ser considerados como diferentes formas de conhecimento relevantes para o entendimento de linguagem natural. Cada estágio contempla um conjunto de propriedades inerentes às palavras, ou seja, as palavras podem ser definidas e vistas sob óticas diferentes em cada estágio permitindo a agregação de mais conhecimento específico.

Allen (1995) apresenta um conjunto mais amplo de estágios ou conjuntos de conhecimentos para o entendimento geral da linguagem natural. Ele realiza a decomposição em cinco estágios como pode ser visto na Figura 1:

- Morfológico: Diz respeito a como as palavras são construídas desde a mais básica unidade de significado chamada morfema.
- Sintático: Compreende o conhecimento de como as palavras podem ser agrupadas para formar sentenças corretas e determinar qual a regra estrutural que cada palavra possui na sentença;
- Semântico: Diz respeito ao significado das palavras e qual a ideia gerada por esses significados quando eles são arranjados em uma sentença. O nível semântico é o estudo do significado dependente do contexto, ou seja, a significação que uma palavra possui dependendo do contexto no qual ela foi usada;
- Pragmático: Refere-se a como as sentenças são usadas em diferentes situações e como esse uso afeta a interpretação da sentença;
- Discurso: Trata do conhecimento de como as sentenças imediatamente precedentes afetam a interpretação da próxima sentença. Essa informação é especialmente importante para a interpretação de pronomes e para a interpretação dos aspectos temporais das informações transmitidas.

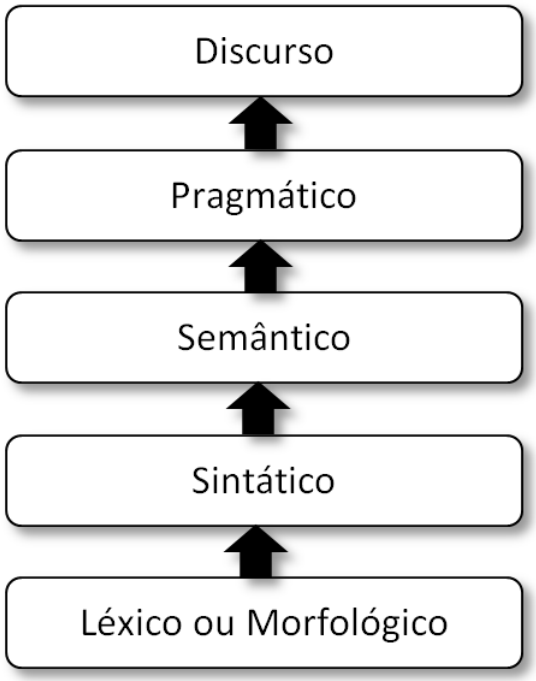

Figura 1 – Os estágios da análise em processamento de linguagem natural

É importante frisar que esses estágios são como degraus de uma escada. Cada estágio acima gera um conjunto de dificuldades maior que o estágio anterior, ou seja, o tratamento computacional de cada estágio é mais complicado. O "subir degraus" faz com que a linguagem tratada esteja mais próxima ao ser humano, e, portanto, mais distante da linguagem computacional.

1.1.1 Análise de Sentimentos

Segundo Liu (2010), de maneira geral, as informações textuais encontradas no mundo podem ser classificadas em dois tipos principais: fatos e opiniões. Fatos são expressões objetivas sobre entidades, pessoas, eventos e suas propriedades, já as opiniões são expressões subjetivas que descrevem o sentimento das pessoas sobre essas entidades, eventos e suas propriedades.

Com a evolução das Redes Sociais Online (RSOs), o número de opiniões emitidas por usuários na Web cresceu intensamente e fez surgir o campo de estudos em Análise de Sentimentos, que visa sumarizar conhecimento, a partir de dados não estruturados, de forma que seja possível para o ser humano extrair informações deste sumário.

A Análise de Sentimentos pode ser encontrada na literatura com diversos nomes: Mineração de Opiniões, Extração de Opiniões, Mineração de Sentimentos, dentre outros. Apesar da grande quantidade de nomes todos eles estão contemplados pelos termos Análise de Sentimentos ou Mineração de Opiniões. Em se tratando de campo de pesquisa, a Análise de Sentimentos é recente tendo seu início por volta dos anos 2000, além disso, é uma

subárea de PLN compartilhando elementos de diversas outras subáreas de pesquisa, como por exemplo: Recuperação de Informações, Mineração de Textos, Aprendizado de Máquina, etc. Uma comprovação para essa afirmação é o aparecimento de trabalhos de Mineração de Opiniões em publicações das áreas correlatas (LIU, 2012).

A Análise de Sentimentos é aplicada sobre qualquer porção de texto de qualquer tamanho e formato, como por exemplo, páginas web, comentários em sites de vendas, *tweets*, posts em blogs, entre outros. E desses textos busca-se extrair e analisar conteúdo subjetivo que representa opiniões de pessoas sobre um alvo (LIU, 2012). Ou ainda conteúdo objetivo que gera opiniões até mesmo sem palavras que possuam sentimentos.

Formalmente, Liu (2010) define opinião como uma quíntupla $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, onde:

- e_i: é o nome de uma entidade;
- a_{ij}: é um aspecto da entidade e_i (opcional);
- s_{ijkl}: é a polaridade do sentimento sobre aspecto a_{ij} que tem como alvo a entidade e_i, emitido por h_k no instante t_l;
- h_k: é o detentor do sentimento (isto é, quem expressou o sentimento), também chamado de fonte de opinião (do inglês: *opinion holder*);
- t_l: é o instante no qual a opinião foi expressa por h_k.

É importante destacar que nem sempre todos os elementos estão presentes em um texto. Basicamente, uma opinião é composta por pelo menos dois elementos que são indispensáveis, o alvo da opinião e o sentimento sobre ele.

O alvo pode ser uma entidade ou um aspecto de uma entidade, um produto, pessoa, organização, marca, evento, entre outros. Com respeito ao sentimento, ele representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. O sentimento expresso pelo autor da opinião define o seu ponto de vista que pode ser representado por alguma escala, revelando, em resumo, um sentimento positivo, negativo ou neutro sobre o alvo da opinião, normalmente denominado de polaridade.

Especificamente, neste trabalho, o alvo das opiniões são produtos e os textos a serem analisados são comentários de clientes sobre estes produtos. A Figura 2 mostra um exemplo de comentário que foi extraído do site *www.buscapé.com.br*.

A partir deste comentário foram extraídas as seguintes tuplas:

- (moto x, geral, positivo, thiago alves, 30/09/2013)

Figura 2 – Exemplo de comentário no buscapé

- (moto x, design, positivo, thiago alves, 30/09/2013)
- (moto x, camera, positivo, thiago alves, 30/09/2013)
- (moto x, armazenamento, negativo, thiago alves, 30/09/2013)

É perceptível que esse formalismo encaixa-se perfeitamente no tratamento de opiniões baseando-se nas suas características. Com relação ao tamanho do texto, é possível distinguir a análise de opiniões em níveis diferentes de granularidade:

- **Documento**: Neste nível, cada texto extraído representa uma opinião e o objetivo deste nível de análise é avaliar o texto como um todo e categorizá-lo com seu respectivo sentimento. A análise a nível de documento não é capaz de capturar as especificidades do texto e se houver outras opiniões no texto analisado todas serão tratadas como uma só.
- **Sentença**: A tarefa neste nível tem o objetivo de analisar as sentenças do texto e verificar se cada sentença é positiva, negativa ou neutra. Tal nível está muito relacionado com a classificação de subjetividade, que busca verificar se um texto é uma opinião, ou seja, expressa o sentimento de alguém por algo (sentença subjetiva), ou se é uma expressão que representa um fato (sentença objetiva).
- **Aspecto ou Característica**: Tanto a análise a nível de documento quanto a nível de sentença não descobrem realmente quais detalhes de um produto as pessoas gostam ou não gostam. O nível de aspecto ou característica oferece uma granularidade mais fina. Ao procurar nas construções linguísticas (documento, parágrafo, sentença), a análise segue a um novo nível buscando diretamente as opiniões, baseando-se na ideia de que uma opinião consiste basicamente de um sentimento (positivo ou negativo) e um alvo. Baseado neste nível de análise, um sumário estruturado de opiniões sobre as entidades e suas características pode ser criado, gerando assim uma estruturação sobre dados que naturalmente são não-estruturados e tais dados podem ser usados para todos os tipos de análise qualitativas e quantitativas.

Dos níveis apresentados, o mais desafiador é o nível de aspecto, pois envolve vários problemas de PLN considerados não resolvidos tais como reconhecimento de entidades nomeadas, resolução de anáfora, escopo de negação e desambiguação de sentidos (LIU, 2010).

Além da divisão por níveis de abstração da análise de sentimentos, as opiniões podem ser divididas em tipos, que definem o quão difícil é a análise destas, pois a forma como as opiniões são expressas influencia diretamente na habilidade de processá-las corretamente.

De modo geral, as opiniões podem ser:

- **Regulares ou Comparativas**: Em opiniões regulares o autor da opinião expressa o seu sentimento, emoção, atitude ou percepção sobre um alvo. Já nas comparativas o sentimento é expresso com base na relação de similaridades ou diferenças entre duas ou mais entidades, ou ainda preferência quanto a algum aspecto compartilhado. Como exemplo, a Figura 3 demostra um comentário que possui ambos os tipos de opiniões. Quando o autor do comentário usa a expressões "*Celular fenomenal*", "*tela de altíssima definição*" e "*Ótima memória*" ele fala diretamente sobre o produto avaliado e suas características, porém, em determinado local no comentário o consumidor apresenta as seguintes descrições: "*Muito melhor que o Iphone 5*" o autor está utilizando na sua argumentação uma comparação entre o aparelho da *Apple* e o aparelho da *Samsung*.

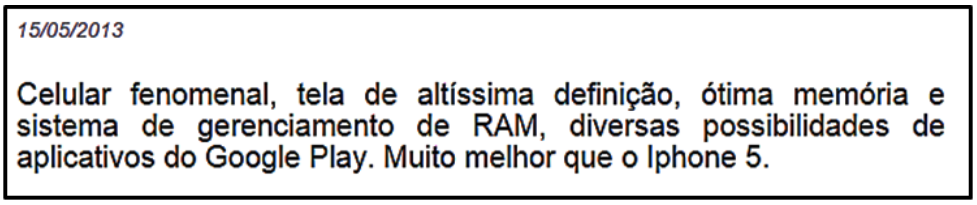

Figura 3 – Um exemplo de opiniões regulares e comparativas

- **Diretas ou Indiretas**: As opiniões podem ainda ser diretas quando referem-se diretamente a uma entidade ou um aspecto da entidade. No caso das indiretas uma opinião é expressa indiretamente sobre uma entidade ou aspecto da entidade baseando-se nos seus efeitos sobre outras entidades. A Figura 4 mostra um comentário com opiniões diretas e também indiretas. O autor do comentário ao argumentar sobre tamanho do aparelho, cita a característica diretamente dando o seu sentimento sobre ela (*Pessoalmente acho o tamanho incômodo*). Quando o mesmo autor trata no início do comentário sobre o aparelho ser ágil e fluido, ele indiretamente está se referindo ao processamento do aparelho, ou um conjunto de características que tornam um aparelho ágil e fluido (Processamento, memória, animações, etc.).

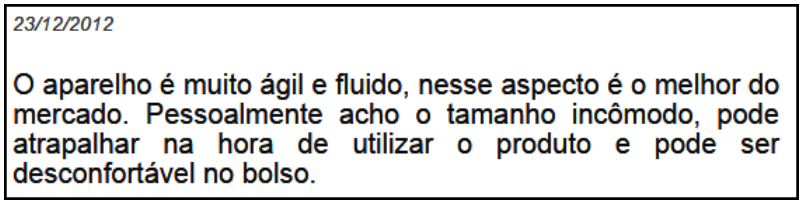

Figura 4 – Um exemplo de opiniões diretas e indiretas

- **Implícitas ou Explícitas**: As opiniões explícitas expressam diretamente o sentimento, enquanto as implícitas sugerem-no indiretamente. No comentário da Figura 5 podem ser observadas estes dois tipos de características. No início do comentário o autor utiliza o termo "*caro*" para tratar implicitamente do preço alto do produto, se o autor utilizasse diretamente a característica do produto, este poderia ser considerado explícito. Outro exemplo de opinião implícita está no final do comentário quando o autor trata da tela do aparelho: "*(...) é que os produtos vendidos no Brasil nunca utilizam a tecnologia Gorilla Glass em suas telas.*". Neste trecho, a referência está na durabilidade da tela do aparelho, dado que a tecnologia *Gorilla Glass* citada no texto, permite uma rigidez e tolerância maior às telas de aparelhos. É perceptível a busca de informações para associar uma dada opinião à característica citada pelo autor, bem como à sua possível orientação semântica.

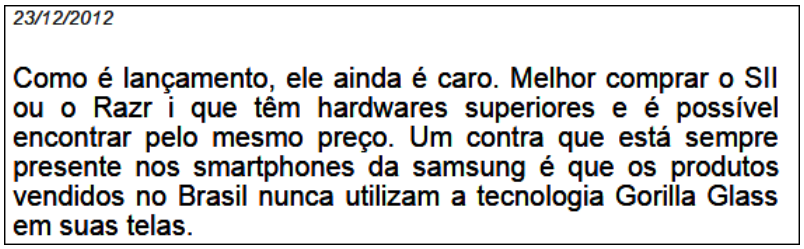

Figura 5 – Um exemplo de opiniões implícitas e explícitas

1.2 Sistemas de Inferência *Fuzzy*

O conceito de conjuntos *fuzzy* tem sido usado na área de Análise de Sentimentos para inferir o grau de positividade ou negatividade de uma opinião (AL-MAIMANI; SALIM; AL-NAAMANY, 2014). Esses conceitos foram introduzidos por (ZADEH, 1965) e referem-se a classes de objetos que não tem uma fronteira estritamente definida, mas, ao invés disso, todos os objetos possuem um grau de pertinência em cada classe. Os conjuntos *fuzzy* são caracterizados por permitir a mudança do grau de pertinência de um objeto de uma classe para outra, suavemente. Esse conceito aliado aos conceitos já estabelecidos na lógica clássica permitiram a construção da Lógica *Fuzzy* (ZADEH, 1975a; ZADEH, 1975b).

A inferência *fuzzy* é um processo de avaliação de entradas (variáveis) com o objetivo de obter conclusões utilizando-se a teoria de conjuntos *fuzzy*, através de regras previamente definidas e das entradas fornecidas. Em outras palavras, um sistema de inferência *fuzzy* é um modelo computacional que utiliza a teoria de conjuntos *fuzzy* e lógica *fuzzy* de forma a lidar com processos de alta complexidade, associados com imprecisões, incertezas e informações qualitativas (MENDEL, 2001).

Os conceitos de regras de *produção fuzzy* e variáveis linguísticas são especialmente importantes para o entendimento dos sistemas de inferência *fuzzy*:

- Uma regra de produção *fuzzy* é basicamente uma regra condicional formada de duas partes principais: *se <antecedente> então <consequente>*. O antecedente é composto por um conjunto de condições que, quando satisfeitas, determinam o processamento do consequente da regra por um mecanismo de inferência *fuzzy*. Já o consequente é formado por um conjunto de ações que serão gerados a partir do disparo da regra.

- Uma variável linguística é uma entidade utilizada para representar de modo impreciso e, portanto, linguístico, um conceito ou uma variável de um dado problema (MARRO et al., 2010). Ela admite como valores apenas expressões linguísticas, como "frio", "muito grande", "aproximadamente", etc. Tais valores contrastam com os valores assumidos por uma variável numérica, que admite apenas valores precisos (ou seja, números). A principal função das variáveis linguísticas é fornecer uma maneira sistemática para uma caracterização aproximada de fenômenos complexos ou mal definidos (TANSCHEIT, 2004). Por exemplo: Na premissa "*Se o nível é baixo...*" *nível* é uma variável linguística e baixo é uma variável *fuzzy*, ou seja, um valor da variável linguística "nível".

Os modelos de inferência *fuzzy* mais utilizados são: o modelo de *Mamdani* (MAMDANI; ASSILIAN, 1975) e o modelo de *Takagi-Sugeno-Kang* (TSK) (TAKAGI; SUGENO, 1985; SUGENO; KANG, 1988). O modelo de *Mamdani* foi um dos primeiros sistemas de controle a serem desenvolvidos tendo por base a teoria de conjuntos *fuzzy* e a lógica *fuzzy*. Já o modelo TSK é similar ao modelo de *Mamdani* em muitos aspectos, como por exemplo, o uso de uma base de regras condicionais de inferência (Observe que os dois modelos seguem o mesmo modelo apresentado na Figura 6). A principal diferença entre os modelos é com relação aos consequentes das regras *fuzzy*. No modelo de *Mamdani*, variáveis linguísticas são utilizadas nos consequentes, enquanto que no modelo TSK são usadas funções polinomiais.

A Figura 6 apresenta o modelo geral de um sistema de inferência *fuzzy*. Os componentes desse esquema geral merecem explicações adicionais (ALMEIDA; EVSUKOFF, 2003; MARRO et al., 2010; TANSCHEIT, 2004):

- Interface de *Fuzzificação*: O modelo de sistema *fuzzy* apresentado, recebe entradas *não-fuzzy*, ou seja, entradas precisas. Em função disso, é necessário efetuar um mapeamento dos dados precisos para os conjuntos *fuzzy* (de entrada) relevantes. A interface de *fuzzificação* é responsável por executar esse mapeamento, preparando a entrada para ser utilizada pelo método de inferência;

- Base de Regras: A base de regras é um conjunto de regras de produção *fuzzy*. Também pode ser conhecida como base de conhecimento *fuzzy*, pois representa o armazenamento de informações do sistema *fuzzy*;

- Inferência: No estágio de inferência ocorrem operações com conjuntos *fuzzy* propriamente ditas. A máquina de inferência recebe valores *fuzzy* provenientes da interface de *fuzzificação*, processa as regras existentes na base de regras e gera a saída para a interface de saída correspondente, a partir da composição de todas as regras disparadas;

- Interface de *Defuzzificação*: O processo de defuzzificação é responsável por obter uma saída numérica a partir dos valores *fuzzy* gerados pela máquina de inferência. Pode ser chamada também de conversão *fuzzy* para escalar, pois transforma informações qualitativas em uma informação quantitativa, sendo um processo de especificação. Diversos métodos podem ser utilizados nessa etapa, os mais utilizados são: o centro de massa e o método da média dos máximos. No caso do primeiro, a saída da inferência é o valor no universo que divide a área sob a curva de pertinência em duas partes iguais. No segundo, a saída precisa é obtida tomando-se a média entre os dois elementos extremos no universo que correspondem aos maiores valores da função de pertinência do consequente.

Assim, quando valores numéricos são aplicados às variáveis de entrada, esses valores são submetidos ao processo de *fuzzificação* para que sejam gerados os valores qualitativos, ou seja, no domínio linguístico das variáveis. Após o processo de *fuzzificação* os valores linguísticos gerados, são aplicados nos antecedentes das regras de produção obtendo assim o valor do consequente para cada regra. Esse procedimento define as regras que foram ativadas com as entradas fornecidas. Todas as regras ativas, contribuirão para a inferência da saída do sistema. Com o resultado de todas as regras ativas, o sistema realiza uma agregação dos resultados. Finalmente, o processo de *defuzzificação* é executado, produzindo um valor numérico como saída do sistema.

Em resumo, o processo de desenvolvimento de um sistema de inferência *fuzzy* normalmente consiste nos seguintes passos:

- Especificar o problema, e definir as variáveis linguísticas;

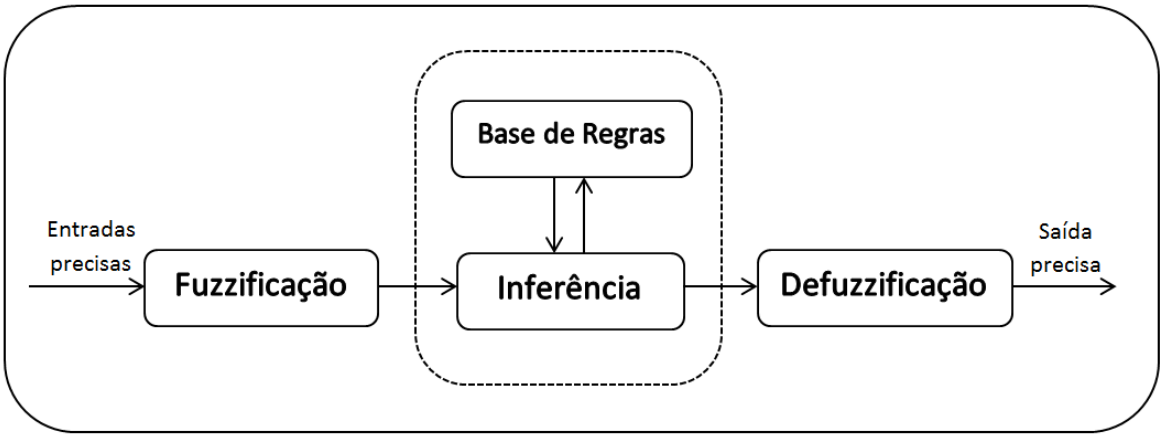

Figura 6 – Esquema geral de um sistema de inferência *fuzzy*

- Determinar os conjuntos *fuzzy*, e as respectivas funções de pertinência para cada variável;
- Definir as regras do sistema de inferência *fuzzy* que devem ser incluídas na base de regras, as quais serão utilizadas para executar as inferências requeridas pelo sistema;
- Avaliar e ajustar o sistema.

Vale frisar que, neste trabalho foi utilizado o modelo de *Mamdani* porque este modelo permite o uso de variáveis linguísticas tanto nas entradas quanto na saída do sistema. Portanto, o processo de modelagem do sistema tornou-se mais intuitivo e com uma melhor interação com o especialista que determinou as regras operacionais do sistema.

1.3 Ferramantas e Recursos

Nesta seção são apresentados as ferramentas e recursos computacionais que foram utilizados no desenvolvimento deste trabalho.

1.3.1 Etiquetadores

Etiquetar morfossintaticamente um texto é atribuir um rótulo ou etiqueta(*tag*) de um conjunto de rótulos (*tagset*) a cada palavra do texto, símbolo de pontuação, palavra estrangeira ou fórmula matemática de acordo com o contexto em que aparecem.

Um *Part-Of-Speech Tagger* (*POS-Tagger*) é um sistema responsável por etiquetar cada item lexical ao analisar um texto dado como entrada, sendo que o conjunto de rótulos são as classes gramaticais pertencentes à língua envolvida na etiquetagem referente a sua categoria morfossintática ou gramatical (substantivo, verbo, adjetivo). Em resumo, o processo de etiquetar consiste, em dada uma sequência de símbolos de um texto e um conjunto de etiquetas, associar a cada símbolo a sua respectiva etiqueta (AIRES, 2000). São

exemplos de etiquetadores: *MXPOST* (RATNAPARKHI, 1996), *TreeTagger* (SCHMID, 1994), *QTag* (MASON, 1997).

O processo de etiquetagem automática basicamente funciona em três etapas: Pré-processamento do texto, classificação gramatical e desambiguação. A etapa de pré-processamento tem o objetivo de preparar o texto de entrada para o formato aceito pelo etiquetador para separação de itens lexicais, por exemplo, o etiquetador *MXPOST* necessita que todos os *tokens* ou itens lexicais estejam separados por um espaço em branco.

Na etapa de classificação gramatical o etiquetador atribui classes gramaticais às palavras. Nesta etapa o etiquetador pode necessitar de informações adicionais para auxiliar o processo, como por exemplo, um léxico ou ainda métodos ou recursos para identificar palavras que não pertencem ao léxico e mesmo com esse conjunto de informações podem haver problemas de ambiguidades. Por existirem esses problemas de ambiguidade, a etapa de desambiguação é necessária e, quando aplicada, normalmente utiliza um conjunto de regras disponíveis para a etiquetagem e o contexto da palavra em questão.

A formação do léxico de palavras e as informações para avaliação do contexto no texto fazem parte do modelo da língua utilizado por cada etiquetador, de acordo com este modelo os etiquetadores podem ser classificados em diferentes classes. De modo geral, os etiquetadores, podem ser **simbólicos ou linguísticos**, quando são baseados por exemplo, em regras, em casos ou em árvores de decisão; ou ainda **probabilísticos ou estatísticos** quando utilizam, por exemplo, representação baseada em *Markov*, árvores de decisão probabilísticas ou distribuições estatísticas de palavras no texto.

Neste trabalho, utilizou- se o etiquetador *TreeTagger*, detalhado em Schmid (1994), que é classificado como um etiquetador probabilístico. Ele foi usado por ser uma ferramenta com boa acurácia de acordo com (GAMALLO; GARCIA, 2013), adaptável a outros idiomas, multiplataforma e ter um conjunto de etiquetas bem definidas para o português. As etiquetas usadas pelo etiquetador *TreeTagger* são descritas no Apêndice A. Destaca-se também que este etiquetador tem sido utilizado em outros trabalhos do grupo de pesquisa em PLN do DC/UFPI

O *TreeTagger* utiliza árvores de decisão probabilísticas e a partir de trigramas (sequências de três palavras consecutivas encontradas em um *corpus*) cria relações entre as classes gramaticais. Para concluir qual a classe gramatical de determinada palavra, é necessário responder afirmativamente ou negativamente a perguntas relativas às palavras que aparecem ao seu redor. À medida em que cada resposta afirmativa é dada, as informações na árvore são conectadas chegando-se a uma resposta: folha da árvore. O etiquetador também possui um léxico que foi criado a partir de uma parte do *corpus Penn Treebank* [1]. Dois milhões de palavras deste *corpus* foram etiquetadas e serviram de treinamento,

[1] http://www.cis.upenn.edu/ treebank/

ou seja, a partir dos dados obtidos neste *corpus*, criam-se regras probabilísticas que são utilizadas na tarefa de etiquetagem de quaisquer outros *corpora*.

É importante destacar que os erros de etiquetagem influenciam os resultados das fases posteriores da análise de sentimentos. Logo, um bom etiquetador é fundamental para a qualidade de uma ferramenta de análise de sentimentos.

1.3.2 Lematização

Em lexicografia uma palavra que é apresentada em sua forma canônica, ou seja, sem flexões de gênero, número ou grau é denominada lema. De forma mais específica, de acordo com Lucca e Nunes (2002) a lematização consiste em reunir todas as ocorrências de uma mesma palavra sob uma única forma, como acontece em um dicionário, em vez de apresentá-las como aparecem em um texto, com variações de gênero, número e grau.

Segundo Galisson e Coste (1983) lematizar é um modo de agrupamento padrão das variações de uma palavra, com a finalidade de simplificar a apresentação e dessa forma facilitar a consulta em conjuntos léxicos em geral, por exemplo, em dicionários. Em se tratando de dicionários a lematização consiste em encontrar um item que represente todas as formas que uma determinada palavra pode tomar. Dessa forma, o infinitivo normalmente é escolhido para simbolizar todas as formas do paradigma verbal (por exemplo: o lema Ter representando as outras flexões: tenho, temos, terei, etc.); e o masculino singular representa todas as formas do paradigma nominal e do paradigma adjetival.

A lematização é especialmente interessante nos momentos em que as flexões das palavras não são importantes, por exemplo, no trabalho de (ANTIQUEIRA, 2007) a lematização é parte importante do processo de agrupamento estatístico de palavras, pois as flexões podem gerar erros no procedimento, causando o aparecimento de palavras indesejadas. Outros trabalhos que fazem grande uso da lematização são os que trabalho na geração, comparação e avaliação de dicionários para línguas naturais (ZANATTA; MIRANDA, 2008; SIMÃO, 2014).

Neste trabalho a lematização foi aplicada nos comentários de usuários, para que estes estivessem prontos para serem comparados com os vocábulos existentes na base do SentiLexPT. As bases do SentiLexPT possuem tanto as formas canônicas das palavras quanto as palavras flexionadas, a lematização simplificou o processo de busca e comparação dos termos existentes nos comentários com a base de referência. Para a extração dos lemas das palavras foi utilizado o *TreeTagger*, pois além da etiquetagem de palavras, ele fornece a forma canônica destas.

1.3.3 Gerador de Parser e Lexer - ANTLR

O ANTLR (PARR, 2013) é um gerador de parser que automatiza a construção de reconhecedores para linguagens de domínios específicos. Ele gera uma gramática combinada que especifica tanto o *parser* quanto as regras léxicas da linguagem. O formato dos *tokens* da linguagem desejada é definido por meio de expressões regulares. Já as regras gramaticais do parser são descritas por meio de uma gramática livre de contexto, na notação EBNF (*Extended Backus Naur Form*).

A Figura 7 apresenta um exemplo de gramática de entrada no formato aceito pelo ANTLR. Na gramática apresentada, podem ser observadas os dois tipos de regras utilizadas pelo ANTLR, as regras léxicas e as regras sintáticas. A diferenciação entre as duas se dá pelos seus nomes, as regras léxicas possuem a letra inicial maiúscula, diferentemente das regras sintáticas que iniciam-se com letra minúscula. Na imagem as regras léxicas são: *DIGIT, PLUS, MINUS, MULT e DIV*, já as regras sintáticas são: *list, term e factor*.

```
grammar OperatorsVer2;

DIGIT : '0'|'1'|'2'|'3'|'4'|'5'|'6'|'7'|'8'|'9';
PLUS : '+';
MINUS: '-';
MULT: '*';
DIV: '/';

list : term ((PLUS|MINUS) term)*;
term : factor ((MULT|DIV) factor)*;
factor : DIGIT;
```

Figura 7 – Exemplo de gramática de entrada do ANTLR (PARR, 2007)

O ANTLR foi amplamente utilizado neste trabalho para a separação de palavras (tratados como *tokens*) em diversos momentos da abordagem. O uso dessa ferramenta evita problemas corriqueiros da língua portuguesa ao se tratar automaticamente com textos, como por exemplo, a acentuação. Para isso cada palavra é modelada por meio de uma regra léxica, que permite seu tratamento adequado. Fez-se uso do ANTLR principalmente para o desenvolvimento de um parser para identificar os padrões linguísticos utilizados como entrada para o Sistema *Fuzzy* apresentado no Capítulo 3.

1.3.4 A Biblioteca *JFuzzyLogic*

A *JFuzzyLogic* (CINGOLANI; ALCALA-FDEZ, 2012) é uma biblioteca *open source*, extensível e independente de plataforma para sistemas *fuzzy* que permite modelar controladores de lógica *fuzzy* (FLC). Um FLC é um sistema de regras baseado em lógica *fuzzy*.

```
FUNCTION_BLOCK tipper

VAR_INPUT
  service, food : REAL;
END_VAR

VAR_OUTPUT
  tip : REAL;
END_VAR

FUZZIFY service
  TERM poor := (0, 1) (4, 0) ;
  TERM good := (1, 0) (4,1) (6,1) (9,0);
  TERM excellent := (6, 0) (9, 1);
END_FUZZIFY

FUZZIFY food
  TERM rancid := (0, 1) (1, 1) (3,0);
  TERM delicious := (7,0) (9,1);
END_FUZZIFY

DEFUZZIFY tip
  TERM cheap := (0,0) (5,1) (10,0);
  TERM average := (10,0) (15,1) (20,0);
  TERM generous := (20,0) (25,1) (30,0);
  METHOD : COG;      // Center of Gravity
END_DEFUZZIFY

RULEBLOCK tipRules
  Rule1:  IF service IS poor OR food IS rancid THEN tip IS cheap;
  Rule2:  IF service IS good THEN tip IS average;
  Rule3:  IF service IS excellent AND food IS delicious THEN tip IS generous;
END_RULEBLOCK

END_FUNCTION_BLOCK
```

Figura 8 – Exemplo do código de uma FCL (CINGOLANI; ALCALA-FDEZ, 2012)

A biblioteca foi desenvolvida seguindo os padrões especificados pela IEC-61131-7 (IEC, 2000), inclusive no que diz respeito aos padrões referentes à Linguagem de Controle *Fuzzy* (FCL, do inglês), com menor esforço e menor curva de aprendizado. De acordo com os desenvolvedores da biblioteca, foram avaliados 25 pacotes de Lógica *Fuzzy*, a maioria do *SourceForge* ou *Google-Code*, que são considerados os mais respeitáveis repositórios de software, e chegaram a conclusão que apenas oito dos pacotes compilaram corretamente e tiveram ampla funcionalidade. E apenas dois deles foram capazes de analisar códigos FCL sob o padrão IEC-61131-7.

Dentre as diversas especificações do padrão para as FCL, a biblioteca permite a implementação de um tipo especial de sistema: o Sistema de Inferência *Fuzzy* (do inglês: FIS). Como exemplo, a Figura 8 mostra um FIS simples utilizando FCL. Este exemplo, que é considerado um típico "*Hello World*" para sistemas *fuzzy*, serve para calcular a gorjeta recebida por garçons em um restaurante. Esse arquivo apresenta locais específicos para declaração de variáveis linguísticas (*var_input* e *var_output*), atribuição de valores para a representação gráfica dos valores *fuzzy* (*fuzzify* e *defuzzify*), para declaração das regras de produção e métodos de fuzzificação e defuzzificação (*defuzzify* e *ruleblock*).

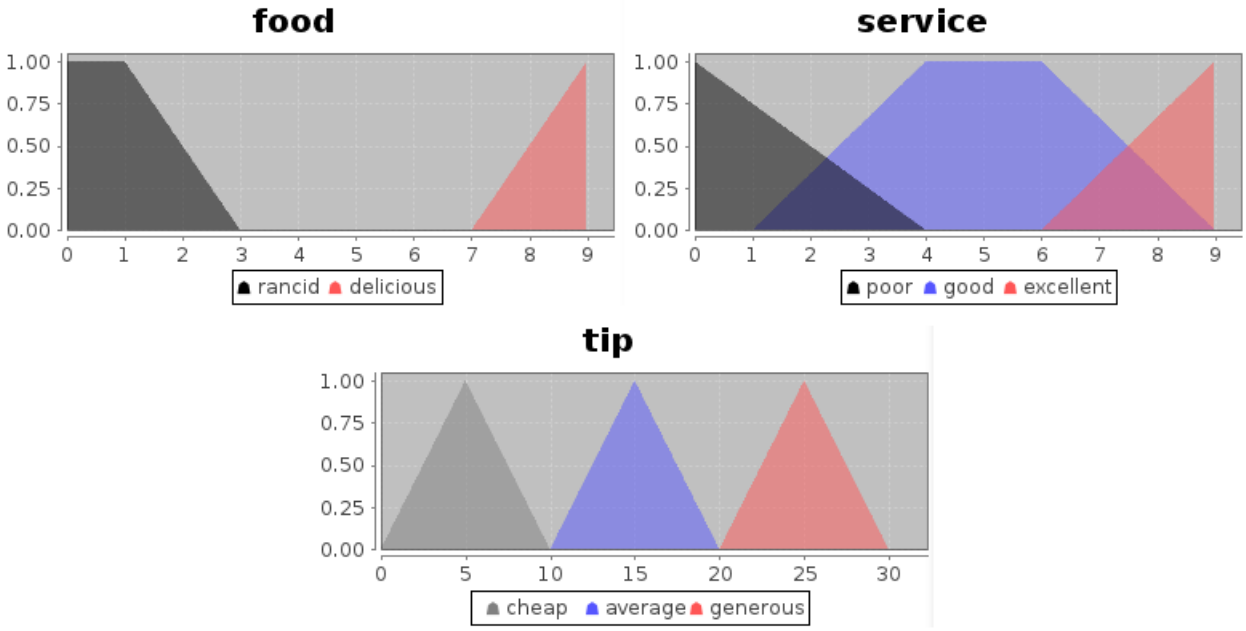

Figura 9 – Um exemplo de gráfico de pertinências gerado pela jFuzzyLogic (CINGOLANI; ALCALA-FDEZ, 2012)

A biblioteca ainda dispõe de diversas variações de funções de pertinência. As principais funções suportadas pela biblioteca são: Funções Triangulares, Trapezoidais, Gaussianas, dentre outras. Outro ponto positivo da ferramenta é a disponibilidade de diversos métodos de defuzzificação, por exemplo: centro de gravidade, centro de área, Máximo à esquerda, dentre outros. Além da disponibilidade de diversas funções de pertinência e métodos de defuzzificação, a biblioteca ainda permite a geração de gráficos explicativos para as variáveis linguísticas do sistema, tornando-se assim uma biblioteca completa e simples para uso na geração e modelagem de sistemas de inferência *fuzzy*.

A Figura 9 apresenta um exemplo da geração de gráficos pela *jFuzzyLogic*. Os valores *fuzzy* utilizados nas variáveis linguísticas desse exemplo são do tipo triangular (Por exemplo, *service*: *poor* e *excelent*) ou trapezoidais (Por exemplo, *food: rancid*), pois são graficamente triângulos ou trapézios, respectivamente. Essa forma de representação é um plano cartesiano no qual considera-se o eixo das ordenadas representando os graus de pertinência e o eixo das abscissas representando os valores, dessa forma, o par ordenado (x, y) indica que o valor x possui um grau de pertinência y. No arquivo FCL (8) os valores *fuzzy* triangulares são representados por dois ou três pares ordenados e os trapezoidais são representados por mais de três pares ordenados.

1.3.5 *Corpus* de Comentários de Consumidores - Buscapé

Embora existam diversos trabalhos na área de análise de sentimentos voltados para o português brasileiro, ainda há falta de recursos úteis para uso da comunidade de PLN. Já para a língua inglesa, por exemplo, existem muitos recursos bem construídos e de fácil

acesso pelos pesquisadores. Devido a essa falta de recursos, mais especificamente, a falta de um *corpus* de comentários de consumidores, foi necessária a coleta e construção de um *corpus*, devidamente etiquetado com informações sobre a polaridade dos comentários.

Assim, em Outubro de 2013, foram coletados 2000 comentários do site Buscapé [2], sendo 1000 positivos e 1000 negativos. O Buscapé foi escolhido, por ser um dos maiores sites de armazenamento de comentários sobre produtos, além de possuir uma forma padronizada de construção do site, permitindo, o uso de ferramentas de coleta automáticas. Os comentários foram coletados da seção de *Smartphones* e celulares do site, com o uso da ferramenta *Crawler JSoup*(HEDLEY, 2010). A estrutura de comentários do Buscapé é composta dos seguintes campos (Ver Figura 10):

1. Nome do autor;
2. Quantidade de Estrelas;
3. Data de publicação;
4. Recomendação do Produto;
5. Título do Comentário;
6. Texto livre do Comentário;
7. Campos de Prós e Contras.

Figura 10 – Comentário completo extraído do buscapé

Alguns pontos acerca do formulário dos comentários merecem atenção especial. Primeiro, o campo de nome do autor permite o acesso ao perfil do autor, dessa forma todos os comentários do autor publicados no site, independente do produto, podem ser extraídos. Em segundo lugar, o campo de recomendação do produto pode ser usado para definir se o comentário é positivo ou negativo no site. O autor do comentário tem a liberdade de

[2] *www.buscape.com.br*

decidir se recomenda ou não um produto, e essa recomendação define o local em que serão publicados os comentários, se na lista dos positivos ou dos negativos. Na coleta realizada, o campo de recomendação foi utilizado para definir os conjuntos de comentários positivos e negativos.

Com relação às estrelas dadas pelos consumidores, a Tabela 1 mostra a distribuição de estrelas de acordo com a polaridade atribuída pelo mecanismo de recomendação do site. É perceptível que existem muitos comentários que receberam quatro ou cinco estrelas, mesmo tendo sido considerados negativos.

Tabela 1 – Distribuição de estrelas

	Polaridade	
Estrelas	**Positivos**	**Negativos**
0	0	9
1	11	127
2	0	226
3	6	367
4	188	162
5	795	109

Embora muitos trabalhos utilizem a quantidade de estrelas como forma de definir a polaridade de comentários, muitos outros advertem sobre o uso de estrelas, pois não é uma métrica perfeita. Ao observar a distribuição da Tabela 1, percebe-se que de alguma forma a divisão realizada pelo Buscapé contém erros, logo, foi necessária uma revisão manual dos comentários para a criação de um *corpus* de teste. Além da revisão manual de polaridades dos comentários, houve a necessidade de outra revisão manual, para inferir a importância do comentário; se excelente, bom, suficiente ou insuficiente. As revisões manuais geraram dois *subcorpus* e os procedimentos realizados são detalhados nas subseções seguintes.

1.3.5.1 *Corpus* Revisado

O procedimento de revisão do *corpus* inicial se deu com a ajuda de três voluntários: uma aluna do curso de letras e dois alunos do mestrado em Ciência da Computação da UFPI. Estes voluntários foram incumbidos da tarefa de ler os 2000 comentários, e decidir se cada comentário é positivo, negativo ou neutro.

Para facilitar a tarefa de revisão manual, os comentários foram armazenados em um banco de dados e foram disponibilizados em uma página WEB pertencente ao domínio da UFPI. A página continha os textos dos comentários e a avaliação inicial (positivo ou negativo) que já era indicada pelo Buscapé. Os estudantes decidiram por manter a avaliação do Buscapé ou modificar a polaridade dos comentários.

Após o trabalho manual, os resultados foram avaliados e, para gerar o *corpus* revisado, somente foram considerados os comentários que tiveram sua polaridade definida unanimemente, ou seja, os três revisores concordaram com a polaridade. A Tabela 2 resume os resultados.

Tabela 2 – Quantidade de comentários resultantes no *corpus* revisado

Positivos	**Negativos**	**Neutros**
982	594	44

Houve uma queda na quantidade de comentários, pois 380 comentários foram considerados irrelevantes ou não obtiveram unanimidade e foram descartados, muitos deles não continham texto entendível, outros somente continham palavras sem sentido, dentre outros motivos. Vale observar também que 44 comentários foram considerados neutros, nesse caso, esses comentários possuem tanto conteúdo considerado positivo quanto negativo.

1.3.5.2 *Corpus* de Referência de Importância

Para avaliar a abordagem descrita nesta dissertação, foi necessário criar um *corpus* que contivesse a avaliação da importância dos comentários de consumidores. Assim, foi extraído do *corpus* revisado, um subconjunto de comentários para serem avaliados por um especialista da área de linguística. Essa atividade foi realizada por um professor da UFPI.

O subconjunto extraído contém 350 comentários, extraídos aleatoriamente do *corpus* revisado, mantendo-se a mesma proporção de comentários por estrela, descrita na Tabela 1 (Para cada valor de estrelas foram selecionados comentários aleatórios de acordo com o total de comentários daquela quantidade de estrelas). Esses 350 comentários foram lidos e avaliados pelo especialista, e para cada comentário, foi atribuído um grau de importância dentre estes: Insuficiente (ISF), Suficiente (SF), Bom (BM) e Excelente (EXC), de acordo com a observação de características sobre o produto abordadas pelo autor e a correção de escrita. A escolha das classes de importância foi feita manualmente. Não poderiam ser criadas muitas classes, pois, aumentaria bastante a quantidade de regras do sistema *fuzzy*, e caso e a escolha fosse por menos classes, o sistema *fuzzy* se tornaria extremamente simples. Portanto, a decisão foi utilizar essa quantidade de classes por acreditar que é um valor equilibrado de classes. A Tabela 3 apresenta o resultado da análise e resume as quantidades de comentários por cada classe analisados.

Tabela 3 – Distribuição das importâncias após a revisão

Grau de Importância	**Positivos**	**Negativos**
Excelente	8	7
Bom	46	49
Suficiente	80	81
Insuficiente	34	45

1.4 Considerações Finais

Neste capítulo foram apresentados os principais conceitos sobre PLN, Análise de Sentimentos e Sistemas *Fuzzy* necessários para o entendimento geral da abordagem proposta. Descreveu-se também as Ferramentas e Recursos utilizados na implementação da abordagem e nos experimentos realizados. Um ponto importante sobre o capítulo é a definição e criação dos *corpora* de referência, que são usados pelos métodos de análise. Tais corpora representam mais recursos à comunidade acadêmica de PLN e mineração de opiniões.

2 Trabalhos Relacionados

Neste capítulo serão apresentados uma breve revisão da literatura, mostrando os trabalhos relacionados a diversos temas, com os quais este trabalho faz uso ou menção. Subdividiu-se em três partes: extração de opiniões, definição de orientação semântica e uso de sistemas *Fuzzy* em mineração de opinião. Para melhor acompanhar a leitura do Capítulo a Figura 11 apresenta visualmente a classificação dos trabalhos apresentados a seguir.

2.1 Extração de Opiniões e Características

Como apresentado na Seção 1.1.1, as opiniões podem ser tratadas em diferentes níveis de granularidade: documento, sentença ou característica. Esta seção apresenta trabalhos voltados exclusivamente para o tratamento de opiniões a nível de características, relacionados a métodos de extração de opiniões e características, por meio de diversas técnicas, por exemplo, padrões sintáticos, medidas de dependência, sintagmas nominais, entre outros.

O primeiro trabalho a ser mencionado é o de Liu, Wu e Yao (2006) que propõem uma metodologia para pesquisar opiniões em comentários envolvendo múltiplos produtos. Os autores identificam todas as expressões relacionadas com o domínio, e então, eles as classificam em dois grupos: características e produtos. A técnica *Pontwise Mutual Information (PMI)* foi usada para calcular o escore de cada expressão candidata de acordo com a diferença de ocorrências entre o *corpus* geral e o *corpus* de domínio específico. Os autores também apresentam um algoritmo para prever a dependência entre características e produtos. Todas as opiniões são indexadas como uma tripla $\langle produto, característica, qualidade \rangle$ e em seguida essas triplas são utilizadas para recuperar opiniões que "casam"com os interesses dos usuários. Além da definição de um novo método de extração de características, o trabalho propõe um método para identificar a dependência entre produtos e características mesmo que a característica seja mencionada em qualquer lugar do texto do comentário. Esse método é baseado na dependência semântica entre as características e o produto. E eles reportaram bons resultados para os métodos propostos.

O trabalho de Hu e Liu (2004) foi um dos pioneiros na área de análise de sentimentos a nível de características. Esse trabalho apresenta uma abordagem para extração de características e opiniões baseando-se principalmente na frequência de termos e na mineração de regras associativas. O sistema proposto pelos autores seleciona os termos frequentes dentre todos os comentários existentes e torna-os características se vierem próximos a adjetivos. Adicionalmente, realiza a extração de possíveis características infrequentes. Caso

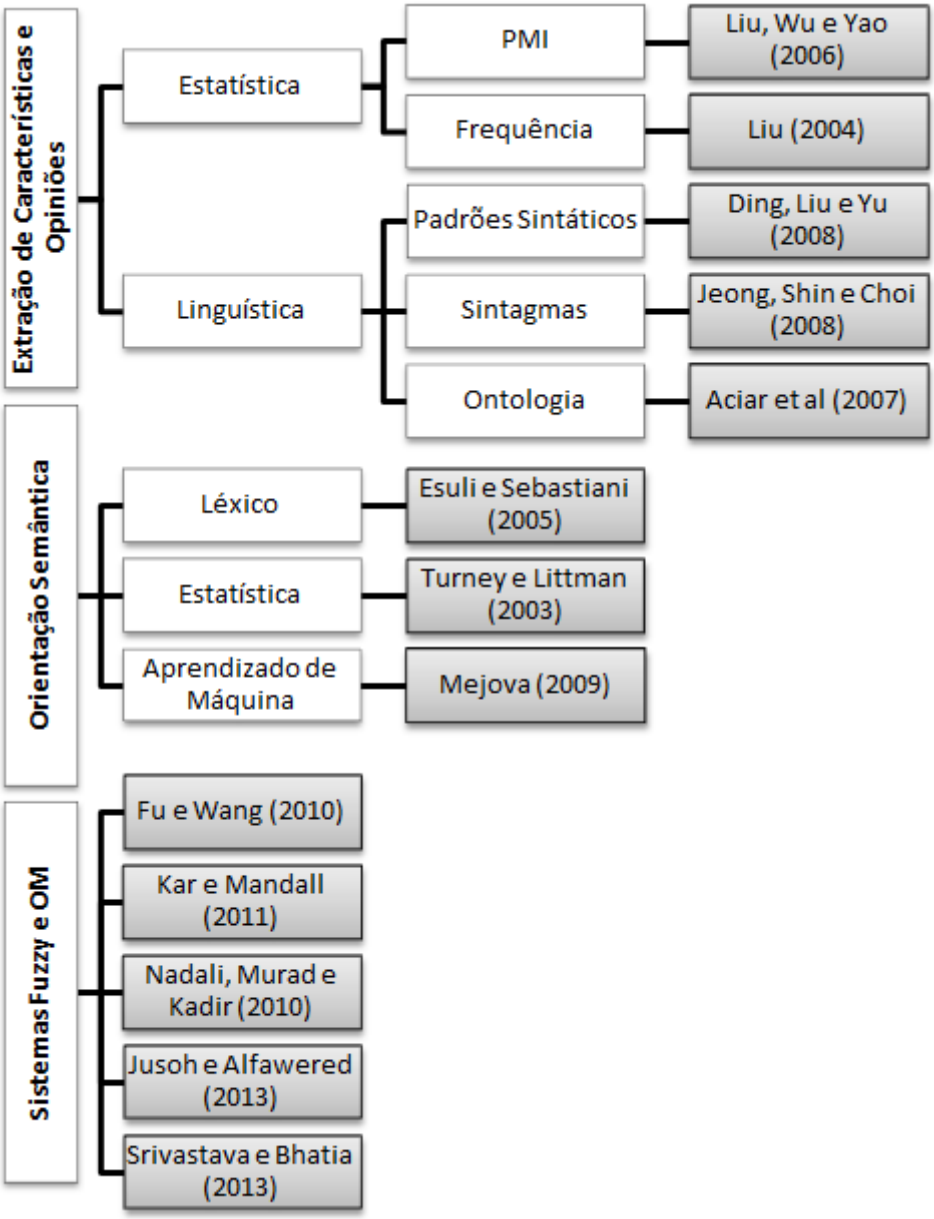

Figura 11 – Classificação de trabalhos apresentados como trabalhos relacionados

não haja uma característica frequente na sentença avaliada, mas se houver uma ou mais palavras opinativas nas proximidades de um substantivo, o substantivo é considerado, portanto, como uma característica não-frequente. Os autores reportaram bons resultados, com precisão (*Precision*) em média 80% e cobertura (*Recall*) média de 72%.

Ding, Liu e Yu (2008) propõem um método de extração de palavras opinativas de comentários escritos em inglês. O método proposto pelos autores utiliza a frequência de termos para gerar uma lista de características contendo tanto características explícitas quanto implícitas. As características explícitas são geradas pela frequência de substantivos encontrados nos comentários, já as implícitas são inferidas de acordo com os verbos e alguns adjetivos. Após gerar a lista de características dos produtos, o sistema busca

fazer a associação das características com as palavras opinativas dependendo da distância entre elas. Dessa forma, o método pode encontrar a orientação semântica das opiniões associadas às características dadas. O trabalho ainda apresenta regras de associação para tratar de negações, expressões adversativas e opiniões dependentes de contexto, que é a principal novidade reportada pelos autores. Eles tratam ainda de qualificadores que são sinônimos ou antônimos de outros, considerando assim, a mesma orientação semântica para os sinônimos ou a orientação inversa para os antônimos. Os autores reportaram uma avaliação empírica com bons resultados e ainda compararam seus resultados com outras ferramentas existentes, o principal ganho da ferramenta sobre as outras analisadas, foi de melhorar a cobertura sem perder precisão. Uma limitação da ferramenta reportada pelos autores é o não tratamento de características sinônimas, porém, prometeram essa funcionalidade como um trabalho futuro.

Outra abordagem para extrair características foi proposta por Aciar et al. (2007) e faz uso de uma ontologia. Os autores do trabalho propõem um sistema de recomendação de produtos. Dessa forma, eles desenvolveram e aplicaram uma ontologia para mapear a qualidade de opiniões e de produtos em um formato próprio para o procedimento de recomendação. Apesar do método funcionar bem semanticamente, uma limitação é a necessidade de manter a ontologia atualizada para resolver o problema da contínua expansão de dados nos comentários. Nessa abordagem a ontologia foi criada manualmente e as atualizações devem ser executadas quando novas características são adicionadas. Uma vantagem desse trabalho é o tratamento de características sinônimas e características implícitas. Cada característica da ontologia possui uma lista de palavras relacionadas. Assim, após o sistema classificar uma sentença, ele busca as palavras dessa sentença nas listas de palavras e assim pode determinar a qual característica essa sentença se refere.

Jeong, Shin e Choi (2011) propõem um sistema para extração e refinamento de características baseando-se em sintagmas nominais e informações semânticas das palavras, denominado FEROM. Os autores indicam que os métodos de extração de características existentes na época, não tinham resultados satisfatórios, e indicaram dois principais problemas para os baixos resultados: A utilização apenas de informações sintáticas das palavras e o tratamento de características com significado similar como diferentes. Durante uma fase de pré-processamento todos as palavras dos comentários são etiquetadas com suas classes gramaticais e assim os sintagmas nominais são identificados. Na mesma etapa de pré-processamento os comentários são tratados por um separador de sentenças, a fim de manter cada característica e suas palavras opinativas em uma mesma frase.

O processo de extração efetiva de características do FEROM se dá pela seleção dos sintagmas nominais, tornando o núcleo desses sintagmas como uma característica candidata. A característica candidata somente é promovida, se houver na mesma sentença um adjetivo, em caso negativo a candidata é descartada. No processo de extração de características,

o FEROM realiza uma conversão das sentenças caso essas sejam negativas, substituindo a expressão negativa por seu antônimo correspondente, por exemplo, "A qualidade da foto não é boa" seria substituída por "A qualidade da foto é ruim". A principal novidade do FEROM é a unificação das características, por exemplo, as palavras foto, imagem, fotografia são consideradas homogêneas (isto é, elas representam a mesma característica). Esse refinamento e mesclagem de características faz uso das relações de sinonímia existentes na *WordNet*, dessa forma a quantidade de características extraídas é menor e mais efetiva que nos métodos anteriores. Os autores realizaram vários experimentos comparando com ferramentas e métodos existentes, os valores médios de *precision* e *recall* foram 90% e 79,3% respectivamente, obtendo melhores resultados que os métodos comparados. Uma limitação reportada pelos autores foi a falta do tratamento de antônimos, mas indicaram esse adendo como trabalhos futuros.

Tendo em mente os trabalhos apresentados nesta seção, a abordagem TOP(X) apresentada nesta dissertação utiliza um procedimento para extração de características e qualificadores que será explicado em detalhes no Capítulo 3. O método proposto é semelhante ao trabalho de Jeong, Shin e Choi (2011), porém, adicionalmente, utilizou-se a estrutura sintática das sentenças para identificar as características e seus respectivos qualificadores. Este método foi escolhido por ser simples e de fácil implementação.

2.2 Orientação Semântica

Na literatura científica, existem várias abordagens para identificar a orientação semântica de opiniões, as quais podem ser baseadas em léxicos de sentimento, técnicas estatísticas e técnicas de aprendizado de máquina. As primeiras são as mais comuns, mas são muito dependentes da qualidade dos léxicos de sentimentos.

A *WordNet* (FELLBAUM, 1998) é o maior e mais conhecido léxico. Atualmente, alguns métodos a utilizam como base para criar outros léxicos mais específicos (ESULI; SEBASTIANI, 2006; KAMPS et al., 2004; GODBOLE; SRINIVASAIAH; SKIENA, 2007). Uma versão estendida desse léxico, o SentiWordNet (ESULI; SEBASTIANI, 2006), foi construído para dar suporte a aplicações de mineração de opinião e classificação de sentimentos. É importante mencionar que a *WordNet* foi criada, primordialmente, para a língua inglesa, mas existe uma versão para o Português brasileiro, chamada *WordNet.BR* (SILVA; OLIVEIRA; MORAES, 2002). Existe também um léxico de sentimentos para o Português de Portugal que é conhecido como *SentiLex-PT* (SILVA; CARVALHO; SARMENTO, 2012), constituído de 7014 *lemmas* e 82347 formas flexionadas. Além desses recursos e ferramentas, diversos outros podem ser encontrados na Linguateca[1]. A Linguateca é um centro de recursos para a língua portuguesa, ela possui uma infraestrutura distribuída

[1] *http://www.linguateca.pt/*

entre vários grandes centros do país, com o objetivo de concentrar serviços, recursos e ferramentas da comunidade de PLN. Dentre os itens disponibilizados pela Linguateca podem ser citados fóruns para distribuição de informações, catálogos de publicações sobre PLN, catálogos de recursos existentes para a língua portuguesa, catálogo de projetos e autores, dentre outros. Sendo também um repositório de ferramentas que necessitam ser distribuídas.

Os métodos de definição de orientação semântica por meio de léxicos de sentimentos são os mais antigos. Kamps et al. (2004) fez uso das relações semânticas existentes na *WordNet*. Eles definiram um grafo demonstrando os adjetivos contidos na intersecção entre uma lista de termos e a *WordNet*, adicionando um *link* entre dois adjetivos sempre que a *WordNet* indicar a presença de uma relação de sinônimos entre eles. No grafo os autores definiram a distância entre duas palavras como sendo o menor caminho entre duas arestas (adjetivos). Além disso, foram definidas duas palavras de referência (*seeds*), "*good*" e "*bad*". A orientação semântica de um termo é calculada de acordo com a sua distância relativa em comparação às duas "*seeds*". O termo será considerado positivo se a distância relativa for um valor positivo, e vice-versa. Dessa forma, a partir de alguns adjetivos é possível definir a orientação semântica de seus sinônimos. Os autores realizaram um experimento com 663 adjetivos e obtiveram uma acurácia de 66,32%.

Com respeito às técnicas estatísticas, o trabalho de Turney e Littman (2003) utilizam métodos para inferir a orientação semântica de uma palavra por meio da medida de associação estatística com dois conjuntos de referência, um com palavras positivas e outro com negativas. Os autores criaram duas instâncias de experimentação, baseando-se em duas medidas estatísticas de associação de palavras: O Pointwise Mutual Information (PMI) e o Latent Semantic Analysis (LSA). O método proposto pelos autores é alicerçado na hipótese de que se uma palavra é positiva (ou negativa) ela deverá estar fortemente associada a palavras positivas (ou negativas). Portanto, os autores criaram um conjunto de palavras positivas independentes do contexto e outro conjunto similar para as palavras negativas. As medidas de associação PMI e LSA foram calculadas entre cada palavra do textos avaliados e os termos de referência de cada conjunto, dessa forma, a orientação semântica das palavras era positiva caso o resultado da associação estatística fosse positivo e vice-versa. Vale ressaltar que o PMI foi calculado por meio de uma ferramenta de busca chamada Altavista. Buscas eram realizadas com o intuito de verificar a associação ou proximidade das palavras avaliadas com os grupos positivos ou negativos. No trabalho foram apresentados diversos experimentos e bons resultados foram reportados pelos autores. Em resumo os experimentos consistiram em testar 3596 palavras das classes adjetivo, advérbio, substantivo e verbos manualmente analisados e divididos em positivos e negativos. Foi reportado que o método atingiu 82,8% de acurácia.

Outra forma para determinar a orientação semântica de opiniões é utilizar técnicas

de aprendizado de máquina. Estas técnicas utilizam um *corpus* de treinamento (documentos rotulados) como uma fonte de conhecimento para encontrar a polaridade das palavras. Diversas técnicas são utilizadas na literatura, as mais conhecidas adotadas na classificação de comentários são: *Naive Bayes*, *Maximun Entropy* e *Support Vector Machine*. Essas técnicas são as que obtiveram maior sucesso na categorização textual (VINODHINI; CHANDRASEKARAN, 2012). Para as técnicas consideradas como supervisionadas, há sempre a necessidade de dois conjuntos de documentos, um para treinamento e outro para teste. O conjunto de treinamento é utilizado para que o classificador automático possa aprender a diferenciar as características dos documentos e o conjunto de teste é usado para validar a performance do classificador automático. Porém, existem poucos conjuntos de referência para treinamento de métodos de orientação semântica, portanto, é importante citar um conjunto de dados rotulado que está disponível nos projetos TREC, CLEF e NTCIR (MEJOVA, 2009), no entanto, eles são destinados para as linguagens europeias (Inglês, Francês, Alemão, Italiano e Espanhol).

É importante ressaltar que as palavras opinativas, na maioria dos trabalhos são da classe gramatical adjetivo, mas verbos também podem ser considerados (LI et al., 2015). Na abordagem proposta nesta dissertação, o método para inferência da orientação semântica das expressões faz uso do léxico de sentimentos *SentiLex-PT*. Destaca-se que, o método de polarização utilizado nos experimentos realizados faz uso tanto de adjetivos quanto de verbos como palavras opinativas.

2.3 Sistemas *Fuzzy* em Mineração de Opiniões

Atualmente na literatura científica, existem poucos trabalhos utilizando Sistemas *Fuzzy* em métodos de mineração de opinião. Fu e Wang (2010) apresentaram um *framework* baseado na teoria de conjuntos *Fuzzy* para classificar sentimentos a nível de sentença para o Chinês. Em virtude da forma de composição das palavras na língua chinesa, os autores realizaram o cálculo da intensidade do sentimento a partir do símbolo mais básico do Chinês: o morfema. Por meio do uso de um *Chi-Quadrado* modificado, eles puderam definir se os morfemas indicavam ou não sentimento. A partir daí, com o uso de palavras com polaridades conhecidas e o *Chi-Quadrado* foram obtidas as polaridades (intensidade do sentimento) dos morfemas. Decorrida a polarização dos morfemas, a tarefa seguinte foi ampliar o método para o nível das palavras e das frases que fazem uso dos morfemas polarizados. Após a determinação da intensidade do sentimento dos morfemas, palavras e sentenças, uma função de pertinência é usada para identificar o grau de pertinência de cada sentença nos conjuntos *fuzzy* definidos. Os autores definiram três conjuntos *fuzzy*: Positivo, Negativo e Neutro para representar as respectivas classes de polaridades de sentimentos. Os autores realizaram experimentos sobre cada granularidade (morfema, palavra ou sentença), o melhor resultado foi para a granularidade de sentença que obteve

cerca de 45% de *f-measure*.

Kar e Mandal (2011) propuseram um sistema de mineração de opiniões nomeado de *Fuzzy Opinion Miner* (FOM) que visa definir a polaridade e intensidade do sentimento de opiniões, por meio do uso de conjuntos *fuzzy*. O esforço inicial dos autores é para realizar uma sucessão de extrações e atribuições de pesos. O primeiro processo de extração é executado para definir a lista de características. A lista é povoada de acordo com a frequência de substantivos encontrados no conjunto de comentários avaliados, são considerados os substantivos com frequência maior que 20% do total de comentários. Além de servir para a criação da lista, as frequências são utilizadas como peso para cada característica considerada, ou seja, os pesos das características são suas respectivas frequências. A seguir, é construída uma lista de adjetivos e advérbios que são geralmente utilizados para qualificar as características de produtos, nomeada de lista de qualidades. O processo de criação dessa lista é manual e, portanto, específica para o domínio. Semelhantemente à lista de características, foram atribuídos pesos para cada palavra da lista de qualidades, porém, dessa vez a atribuição foi manual. Por exemplo, "bom" = 0.6, "muito bom" = 0.7746, "não adequado" = 0.4, e assim por diante.

Posteriormente, os autores definiram padrões sintáticos para extração de frases opinativas, que são sequencias de palavras que denotam sentimentos das classes de adjetivos, substantivos, advérbios e verbos. Nos padrões definidos, somente foram consideradas as palavras pertencentes às listas de características e qualidades. Por fim foram definidas as funções para calcular a intensidade efetiva das frases opinativas extraídas. Nessa etapa são considerados elementos intensificadores, atenuadores e inversores de sentimentos. O valor final de uma frase opinativa é o produto $A * B$, sendo A o peso da característica qualificada, e B o resultado do calculo de intensidade. Para avaliar a proposta os autores consideraram o *ranking* de sete câmeras fotográficas em sites de venda, em termos de notas. Utilizando o escore gerado pelo FOM, eles conseguiram replicar o *ranking* dos produtos.

O trabalho de Nadali, Murad e Kadir (2010) propôs um sistema *Fuzzy* que executa a classificação de sentimentos de opiniões de consumidores. As opiniões foram classificados em várias categorias (isto é, muito positivo ou negativo, moderadamente positivo ou negativo, fracamente positivo ou negativo e muito fracamente positivo ou negativo). Eles utilizaram adjetivos, advérbios, verbos e substantivos como palavras opinativas. Adicionalmente, o sistema *fuzzy* modelado pelos autores trata essas quatro classes como variáveis de entrada. Portanto, para cada palavra dessas classes foram atribuídos graus especiais. Por exemplo, "excelente" = 6, "bom" = 3, "gostar" = 4, "muito" = 5, dentre outros. Os valores foram definidos por especialistas humanos. Os autores utilizaram uma função de pertinência triangular com três variáveis linguísticas que são *baixo*, *moderado* e *alto*, para as variáveis de entrada. As fronteiras para estes conjuntos foram também definidas por especialistas humanos. Baseados nestes conjuntos, algumas regras *fuzzy* foram desenvolvidas para lidar

com cada caso e, consequentemente, serem utilizadas no processo de *deffuzificação* das saídas. Tendo por base o modelo de *Mamdani*, os autores utilizaram *center of gravity* como operador de deffuzificação. No trabalho, os autores não reportaram quaisquer resultados.

No trabalho de Jusoh e Alfawareh (2013), os autores propuseram o uso de conjuntos *fuzzy* e um léxico de sentimentos com valores *fuzzy* para definir o grau de polaridade (positivo ou negativo) de opiniões. Eles criaram manualmente o léxico de sentimentos utilizando somente adjetivos e advérbios como palavras opinativas (nomeadas por eles de *SenWords*). Para cada palavra do léxico, eles atribuíram manualmente um grau de positividade ou de negatividade. O *framework* proposto busca *SenWords* em comentários de consumidores. Assim, utilizando os valores atribuídos às *SenWords*, ele calcula o grau de positividade ou negatividade final do comentário, por meio do operador *fuzzy MAX*, ou seja, o grau de positividade ou negatividade final é o maior valor entre as *SenWords* existentes no comentário. Os autores conduziram um pequeno experimento com os comentários de consumidores sobre hotéis, apresentando graficamente o grau geral do sentimento dos consumidores. No entanto, eles também não reportaram resultados.

Srivastava e Bhatia (2013) descreveram um sistema de inferência *Fuzzy* com o objetivo de quantificar uniformemente a intensidade de expressões subjetivas quando modificadores adverbiais alteram a força da opinião expressa. Eles utilizaram o léxico de sentimentos SENTIWORDNET 3.0 para construir um léxico próprio de palavras opinativas. O SENTIWORDNET possui três valores referentes à positividade (P), negatividade (N) e objetividade (O) das palavras e, utilizando essas informações, eles coletaram todas as palavras do SENTIWORDNET que possuíam os valores de P e N maiores que zero. Após a coleta, eles executaram uma normalização para gerar apenas um único valor de subjetividade no intervalo [0,1], sendo que quanto mais próximo de zero mais negativo é o termo. Com base nesses valores eles modelaram dois conjuntos *fuzzy*, um para as palavras positiva e outro para as negativas. A principal novidade do trabalho é o tratamento da modificação da intensidade de expressões subjetivas na presença de *n-gramas* de advérbios, por meio do uso de limitantes *fuzzy* (*Fuzzy Hedges*). Para isso, eles definiram classes de advérbios de acordo com as suas respectivas capacidades de modificação (Normal, Forte, Fraco, Minimizador e Negação), esses são os limitantes linguísticos do sistema *fuzzy*. Para cada limitante linguístico eles propuseram uma função de modificação, essa função altera o valor da intensidade do sentimento da palavra opinativa. Além disso, os limitantes possuem a capacidade de composição, ou seja, cada componente dos *n-gramas* realiza uma modificação no valor da intensidade do sentimento. Por exemplo, a sentença "*screen is beautiful*" não possui modificador adverbial, a sentença "*screen is very beautiful*" é modificada por um unigrama adverbial e a sentença "*screen is very very beautiful*" é modificada por um bigrama adverbial. A intensidade final do sentimento expresso pelas sentenças, é definido pela aplicação dos limitantes *fuzzy*. No exemplo dado, o valor de intensidade final das sentenças é, respectivamente, 0.4, 0.632 e 0.796. Eles

conduziram um experimento sobre 50 sentenças com suas intensidades de sentimento atribuídas manualmente por cinco pesquisadores. O erro médio produzido no experimento ao comparar a amostra manual com os valores gerados pelo sistema foi menor que 0.0118, valor esse que os autores consideraram muito pequeno e facilmente aceitável. Entretanto, eles indicaram que algumas técnicas para minimização de erros devem ser utilizadas futuramente, por exemplo a técnica de *back propagation*.

Todos os trabalhos apresentados nesta seção utilizam os conceitos de lógica *fuzzy* ou aplicam sistemas *fuzzy* para estimar a força de opiniões emitidas por consumidores. Tal uso é diferente para o sistema *fuzzy* proposto nesta Dissertação, pois ele fora modelado de modo a inferir da importância dos comentários avaliados. Dessa forma, a principal vantagem da abordagem proposta é permitir classificar (ordenar) e, assim, definir os TOP(X) comentários mais significantes, reduzindo, portanto, a complexidade da tarefa de avaliar os inúmeros comentários de produtos e serviços, disponíveis em web sites e nas redes sociais.

2.4 Considerações Finais

Este capítulo apresentou alguns trabalhos relacionados aos três principais desdobramentos desta dissertação: extração de opiniões e características, definição de orientação semântica e sistemas *fuzzy* na mineração de opiniões. É importante ressaltar que existem muitos outros trabalhos na literatura sobre os temas, um dos objetivos dessa seção foi apresentar alguns trabalhos que merecem destaque nas áreas listadas. Na literatura existem diversas pesquisas (*surveys*) sobre essas áreas, com destaque para os trabalhos de (TANG; TAN; CHENG, 2009; CONRADO et al., 2014; SEERAT; AZAM, 2012; VINODHINI; CHANDRASEKARAN, 2012; AL-MAIMANI; SALIM; AL-NAAMANY, 2014).

Para extração de opiniões e características, a abordagem TOP(X) utiliza um procedimento baseado na estrutura sintática das sentenças e em sintagmas para definir as orações que possuem predicativos do sujeito. Esses métodos foram escolhidos por simplicidade e facilidade de implementação.

Já em relação à inferência da orientação semântica das expressões, este trabalho faz uso do léxico de sentimentos *SentiLex-PT*. Destaca-se que, o método de polarização utilizado nos experimentos realizados faz uso tanto de adjetivos quanto de verbos como palavras opinativas, além de permitir a associação de advérbios como modificadores diretos das palavras opinativas.

Com relação aos sistemas *Fuzzy* em mineração de opiniões, o sistema *Fuzzy* proposto foi modelado para inferir importância de comentários avaliados. Portanto, pode-se ranquear e definir os TOP(X) comentários mais significantes, reduzindo, assim, a complexidade da tarefa de avaliar os inúmeros comentários de produtos e serviços.

Vale observar que o principal objetivo desta dissertação é apresentar a abordagem TOP(X), que busca diminuir o esforço computacional de métodos de análise de sentimento reduzindo a quantidade de opiniões avaliadas por esses métodos. Adicionalmente pode, ainda, facilitar a análise de opiniões por consumidores, também por meio da leitura de poucos comentários importantes. Com relação aos procedimentos computacionais, ao definir os comentários mais importantes os métodos de análise de sentimentos poderão avaliar somente os comentários considerados mais importantes (cerca de 10%), ao invés de buscar e avaliar milhões de comentários na Web e ainda cometendo menos erros e problemas em comparação à avaliação de todos os possíveis comentários. Vale ressaltar que a redução de esforço computacional será sobre o método de análise de sentimentos, nada se pode afirmar ainda sobre o custo computacional global. Já com relação aos consumidores que preferem ler os comentários quando desejarem informações sobre seus produtos, não perderão tempo lendo grandes quantidades de comentários. Esses consumidores poderão ler, por exemplo, os TOP(10) comentários e deles retirar uma decisão segura sobre o produto desejado.

3 Abordagem TOP(X)

Este capítulo tem como objetivo descrever a abordagem TOP(X) desenvolvida para inferir a importância de comentários de usuários Web sobre produtos e serviços. Inicialmente, apresenta-se uma visão geral do problema de pesquisa atacado. Em seguida, discute-se as três variáveis de entrada e a variável de saída que são consideradas no sistema de inferência *fuzzy* proposto na abordagem TOP(X). No final, algumas considerações sobre o modelo proposto são analisadas.

3.1 Visão Geral do Problema de Pesquisa

O comércio eletrônico (*E-Commerce*) é um tipo de transação comercial em franco crescimento na *Web*, principalmente as transações do tipo *Business to Consumer* (B2C) que são voltadas diretamente para o consumidor final. Segundo uma pesquisa encomendada pelo *PayPal* junto à empresa *BigData Corp*, analisando dados de até 1º de Março de 2015, existiam 450 mil sites de *E-commerce* ativos no Brasil [1].

Em relação ao mercado norte-americano, segundo uma pesquisa conduzida pela *Dimensional Research*, realizada no início de 2013, 1046 pessoas completaram um *survey online* sobre suas experiências como consumidores. Essa pesquisa indicou que 88% dos consumidores tiveram suas decisões de compra influenciadas pela leitura de comentários sobre produtos. Em outra pesquisa realizada pelo blog *PeopleClaim*, sobre 2445 pessoas, denominada de *The review of reviews*[2] indicou que 82% desses consumidores consideram os comentários gerados por usuários na Web "extremamente valiosos" ou "valiosos". Além disso, na mesma pesquisa, 70% dos consumidores consultam comentários ou avaliações antes de realizar alguma compra. Outra informação importante dessa pesquisa é que 63% dos consumidores, quando precisam comprar produtos *online*, preferem realizar a compra de um *site* que possui avaliações de produtos e comentários de outros usuários.

Sabendo da importância das avaliações e comentários de produtos, a maioria dos *sites* disponibilizam espaços para que usuários possam comentar e avaliar os produtos adquiridos (exemplo: *americanas.com.br, submarino.com.br, pontofrio.com.br*, etc.). Além de sites que efetivamente são voltados para vendas, existem outros que funcionam também como agregadores de opiniões permitindo que usuários comentem sobre quaisquer produtos cadastrados, centralizando, assim, boa parte das revisões de clientes sobre produtos comprados (exemplos: *buscape.com.br, bondfaro.com.br, zoom.com.br*, etc.).

[1] http://exame.abril.com.br/pme/noticias/o-perfil-do-varejo-online-no-brasil
[2] http://www.peopleclaim.com/blog/index.php/the-review-of-ratings/

Apesar do objetivo desses sites ser permitir ao usuário emitir sua opinião sobre os produtos, eles diferem na forma de capturar a informação. Os principais formatos de campos de comentários na web são:

1. Prós e Contras: Existem locais separados para o usuário descrever os comentários positivos e negativos.
2. Prós, Contras e Revisão detalhada: Além do usuário descrever os prós e contras separadamente, é possível também escrever uma revisão detalhada sobre a sua experiência em geral do produto avaliado.
3. Formato livre: O usuário pode escrever livremente, isto é, não há separação entre Prós e Contras.

Esses formatos são utilizados como objeto de pesquisa pelos diferentes métodos de análise de sentimento de produtos. Porém, a maioria dos trabalhos considera principalmente os formatos que possuem textos livres (itens 2 e 3 da classificação anterior). Vale ressaltar que embora a maneira de expor Prós e Contras seja útil, principalmente para definir as características dos produtos, não é a forma mais natural usada pelo ser humano em comparação a um texto livre. Além disso, opiniões explicando a experiência do cliente sobre um produto permitem análises mais detalhadas sobre o real sentimento que ele pretende transmitir.

Diante do exposto, os principais trabalhos da área concentram-se em analisar as opiniões de clientes sobre produtos, com o intuito de definir a orientação semântica dessas opiniões. Alguns desses trabalhos focam em definir a polaridade (positivo/negativo/neutro) das opiniões, outros buscam ir além da polaridade e definem métricas para tratar a orientação semântica como algo para permitir níveis de positividade ou negatividade, ou seja, consideram que a orientação semântica de uma opinião está em algum ponto entre o positivo e o negativo, no intervalo $[-1, +1]$.

De fato, enquanto os principais trabalhos da área têm extensivamente analisado e classificado sentimentos em opiniões online (GHOSE; IPEIROTIS, 2006) e explorado métodos automáticos para obter conjuntos de atributos, orientação semântica, além de outras características das opiniões, eles não examinam a importância ou utilidade dos comentários, ou seja, o impacto que os comentários causam sobre o resultado final da análise e sobre os consumidores. É importante também frisar que a importância ou utilidade deve ser tratada como uma métrica para definir a qualidade de um determinado comentário e é diretamente proporcional à quantidade de informações que o comentário proporciona ao leitor ou método de análise de sentimento.

Para calcular ou inferir a importância dos comentários de clientes na abordagem TOP(X), foram extraídas informações que, a princípio, são relevantes para tornar um

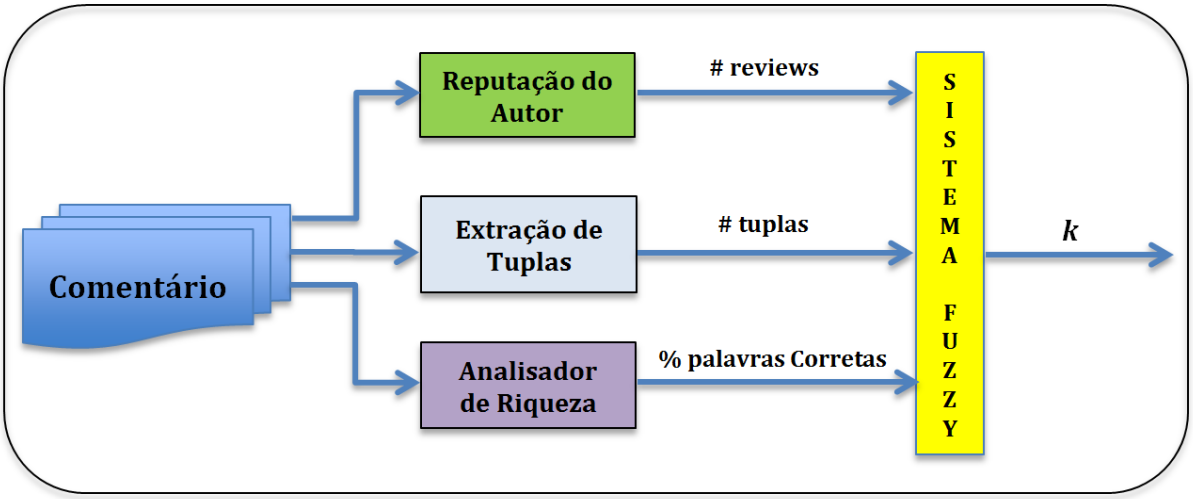

Figura 12 – Estrutura geral do sistema *fuzzy* proposto

comentário mais importante que outro. Vale observar que o foco deste trabalho é propor uma abordagem para inferir a importância das opiniões, logo, outras informações podem ser utilizadas, bem como, outras métricas. A abordagem TOP(X) pode ser usada como modelo para que outros métodos possam ser estudados e propostos.

As variáveis escolhidas para calcular a importância das opiniões neste trabalho foram: reputação do autor, quantidade de padrões linguísticos e o percentual de palavras corretamente escritas nos comentários, o qual será chamado de riqueza de vocabulário. A estrutura geral da abordagem é mostrada na Figura 12. Vale ressaltar que as variáveis foram escolhidas de forma empírica e acredita-se que sejam as principais variáveis para definir a importância de um comentário na Web. A escolha dessas variáveis parte do pressuposto que um comentário escrito corretamente, com grande cobertura das características e realizado por um autor confiável é bastante informativo e, portanto, muito importante. O valor de k refere-se à saída do sistema *fuzzy* e indica o grau de importância do comentário no intervalo [0..10] de números reais. Mais detalhes sobre as variáveis são dados nas seções seguintes.

3.2 Variáveis de Entrada

3.2.1 Reputação do Autor

Definir uma métrica para calcular a reputação de uma pessoa não é plenamente aceito por pesquisadores da área de Humanas, e é motivo de conflitos em fóruns da comunidade de Interação Humano-Computador (IHC). Pesquisas sobre reputação de autor surgiram para tentar resolver o problema da grande quantidade de *spams* na Internet. Normalmente, estes *spams* são gerados por pessoas contratadas ou empresas que estão interessados em aumentar suas vendas e credibilidade. Os trabalhos (JINDAL; LIU, 2008)

e (XU, 2013) objetivam detectar *spams* em comentários acerca de produtos na Internet. Dentre as técnicas mais utilizadas para a detecção de *spams*, estão a análise do próprio texto escrito e a análise do perfil dos autores dos comentários. Portanto, a reputação do autor tem relevância para estimar tanto a validade quanto a importância dos comentários na Internet. De maneira geral, acredita-se que as opiniões de especialistas sobre questões pertencentes a suas áreas de pesquisa têm um peso maior em comparação a pessoas que não têm o mesmo grau de especialização.

A hipótese defendida ao utilizar essa variável, é que os autores que estão frequentemente escrevendo opiniões têm melhor reputação comparando-se a autores ocasionais. Assim, para este trabalho, o valor de reputação dos autores foi calculado pela observação de todos os perfis de autores do *corpus* de referência. Para cada autor foi recuperada a quantidade total de comentários existentes no site e não apenas os comentários existentes no *corpus*. Assim, de acordo com essa métrica, quanto mais comentários o autor emitir, maior será a sua reputação.

Vale destacar que alguns filtros foram utilizados no processo de extração total de comentários de cada autor, por exemplo, não foram considerados os comentários em branco e nem comentários repetidos. Por fim, tem-se a convicção de que a reputação do autor deve ser melhor estudada, pois ela representa uma variável importante para o modelo proposto pela abordagem TOP(X). Destaca-se que existe atualmente um trabalho de mestrado sendo desenvolvido no grupo de pesquisa em PLN do PPGCC/UFPI com o objetivo de explorar minuciosamente as informações dos autores de comentários. O grande desafio desta pesquisa é conectar o perfil do autor de um comentário em um site aos perfis do mesmo autor em outros sites e nas diversas redes sociais existentes.

3.2.2 Quantidade de Tuplas

Em revisões sobre produtos e serviços é comum encontrar características citadas pelos autores nas proximidades das suas respectivas qualidades. Muitas pesquisas foram realizadas para a extração de características e suas qualidades, tais como (LIU; WU; YAO, 2006; ACIAR et al., 2007; PENG; SHIH, 2010). A identificação de tuplas $\langle característica, palavra\ opinativa \rangle$ torna-se relevante para o resultado final do processo de análise de sentimento. Assim, neste trabalho, a hipótese defendida ao utilizar essa variável, é que quanto mais características e qualidades forem citadas pelos autores, mais informativo é o comentário e mais importante.

A utilização da quantidade de tuplas justifica-se pela diferença de profundidade da análise entre os consumidores. Alguns consumidores procuram efetuar uma revisão mais completa sobre o produto adquirido, explicando diversas características do produto e as qualificando. Outros consumidores, no entanto, focam em apenas citar e qualificar as características que não foram plenamente de seu agrado ou vice e versa. Considerando o fato

de que, quanto mais características forem citadas pelos autores, maior será a probabilidade de satisfazer uma dúvida de outro consumidor, um comentário mais explicativo certamente será mais útil para uma quantidade maior de pessoas e, portanto, será mais importante para os métodos de análise.

Para a extração das tuplas na abordagem TOP(X), foi desenvolvido um fluxo de processos, como mostra a Figura 13.

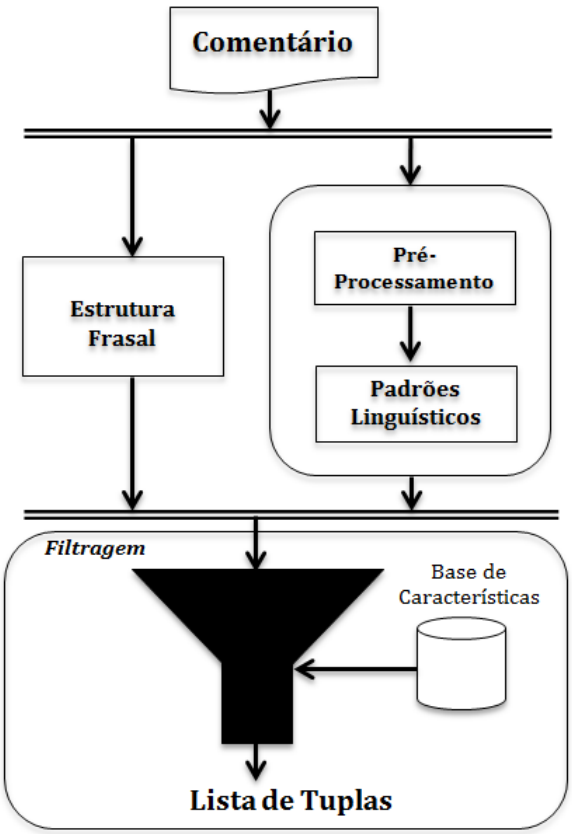

Figura 13 – Fluxo do processo de extração de tuplas

No primeiro processo os comentários são analisados de acordo com a estrutura frasal das sentenças. Nas sentenças com a estrutura "sujeito + verbo de ligação + predicativo do sujeito", o núcleo do sujeito define a característica e o núcleo do predicativo indica a palavra de qualidade. Por exemplo, na sentença: "*A tela é muito boa*", a tupla extraída é $\langle Tela, muito\ boa \rangle$. Nesse procedimento, nenhum pré-processamento do texto foi realizado.

No processo seguinte os comentários são analisados novamente, porém, dessa vez, de acordo com alguns padrões linguísticos pré definidos. Os padrões utilizados são baseados nos especificados por Turney (2002) e em algumas extensões definidas por Kar e Mandal (2011). A Tabela 4 mostra os seis padrões que foram utilizados na abordagem TOP(X) e a Tabela 5 apresenta alguns exemplos de extrações por meio dos padrões linguísticos. Nesse procedimento, uma rotina de pré-processamento foi executada para remover palavras de classes não desejadas (artigo, pronome, número e conjunção), pois essas classes não

apresentam informações relevantes para esse procedimento. Além dessas classes, todos os símbolos de pontuação e dígitos foram excluídos. Um fato importante é que os dois primeiros processos podem ser executados paralelamente sem prejuízos para o resultado.

Tabela 4 – Padrões linguísticos utilizados para extração de tuplas

Padrões Linguísticos
$ADJ\ SUBS\ (PREP?\ SUBS)*$
$ADV\ ADV?\ ADJ\ (SUBS\ (PREP?\ SUBS)*)?$
$SUBS\ (PREP?\ SUBS)*\ ADJ\ ADV\ ADV?$
$SUBS\ (PREP?\ SUBS)*\ ADV?\ ADJ+$
$ADV\ VERB$
$VERB\ ADV$

Tabela 5 – Exemplos da extração de tuplas por meio dos padrões linguísticos

Exemplos
Linda tela!
Bateria dura bastante!
Processador muito lento!
Nunca funcionou!
Quebrou rápido!

O último processo consiste na filtragem das tuplas encontradas. Para essa atividade, foi necessário um trabalho manual para extrair todas as características dos produtos analisados, o qual gerou uma base com 40 características sendo o recurso chave para a filtragem realizada. Neste passo são consideradas somente as tuplas que possuem característica pertencente à **base de características**, todas as outras são excluídas. Destaca-se que este recurso torna a abordagem dependente do domínio dos produtos analisados.

3.2.3 Riqueza de Vocabulário

Em qualquer sociedade, um texto claro, compreensível e bem escrito não pode conter erros ortográficos e gramaticais. Apesar do caráter informal de comentários escritos na web e em redes sociais, inclusive com a definição de dicionários de "internetês", palavras escritas incorretamente podem ser consideradas ruídos, em se tratando de análise de sentimentos de comentários. Autores como Tumitan e Becker (2013) reconhecem que as palavras incorretas podem atrapalhar os resultados e uma verificação ortográfica nesses casos torna-se necessária. Paltoglou e Giachanou (2014) afirmam que uma parte significante das descrições textuais nas mídias sociais contém linguagem não-padrão, incluindo palavras escritas incorretamente e outros problemas, tais como abreviações, substituições fonéticas e

emoticons. Estes problemas causam erros no processo de tokenização, na definição das classes gramaticais, no reconhecimento de entidades nomeadas, assim, afetando negativamente os resultados das análises. Algumas pesquisas sugerem uma etapa de pré-processamento para traduzir do "internetês" para a língua natural a ser analisada.

Na abordagem TOP(X), a hipótese que define o uso da variável de Riqueza de Vocabulário é: quanto mais correto um comentário for escrito, mais útil será a opinião do autor. Assim, um comentário emitido com muitos erros ortográficos, será considerado menos importante do que outro sem erros ou com poucos erros de escrita.

Em uma implementação inicial da abordagem TOP(X), não foi considerado nenhum dicionário para o "internetês" e o percentual de palavras escritas corretamente foi utilizado como valor para essa variável. No entanto, é possível usar outras métricas para avaliar a riqueza do vocabulário, por exemplo, o tamanho do vocabulário, o número de "hapax legomena" (palavras que ocorrem apenas uma vez em um dado *corpus*), entre outras.

Finalmente, para realizar o cálculo do percentual de palavras ortograficamente corretas, foi utilizado a ferramenta *Wiktionary* [3] para a língua portuguesa.

3.3 Sistema *Fuzzy*

Como visto na seção 1.2 (Sistemas de Inferência *Fuzzy*), a abordagem TOP(X) faz uso de um Sistema de Inferência *Fuzzy* do tipo *Mamdani* para inferir a importância de comentários sobre produtos ou serviços. De acordo com o modelo, após a definição das variáveis de entrada deve-se realizar a inferência da variável de saída, que na abordagem TOP(X) será usada para inferir o grau de importância de um comentário. É importante ressaltar que no modelo de *Mamdani* a definição da saída do sistema é apresentada por meio variáveis linguísticas, o que torna a modelagem do sistema mais intuitiva, pois é realizada em um nível de abstração mais próximo do ser humano. Dessa forma, o especialista de domínio pode definir manualmente o grau de importância de um comentário, mesmo sem conhecimentos aprofundados de sistemas *Fuzzy*. O uso de sistemas *fuzzy* para este trabalho é justificado pelo alto grau de incerteza existente na definição de importância de um comentário. Dessa forma, trabalhar com valores linguísticos é mais intuitivo que trabalhar com valores absolutos. Os sistemas *fuzzy* oferecem um arcabouço de facilidades para automatizar o processo de inferência de importância dos comentários.

O processo de modelagem partiu da definição de valores linguísticos para as variáveis de entrada e saída. Assim, para cada variável de entrada, foram associados três valores linguísticos: **Baixo** (B), **Médio** (M) e **Alto** (A). Foram determinados três valores linguísticos pois, considera-se um valor suficiente para a representação das variáveis. Além disso, essa quantidade não torna a modelagem do sistema muito complexa pela

[3] disponível em: http://pt.wiktionary.org.

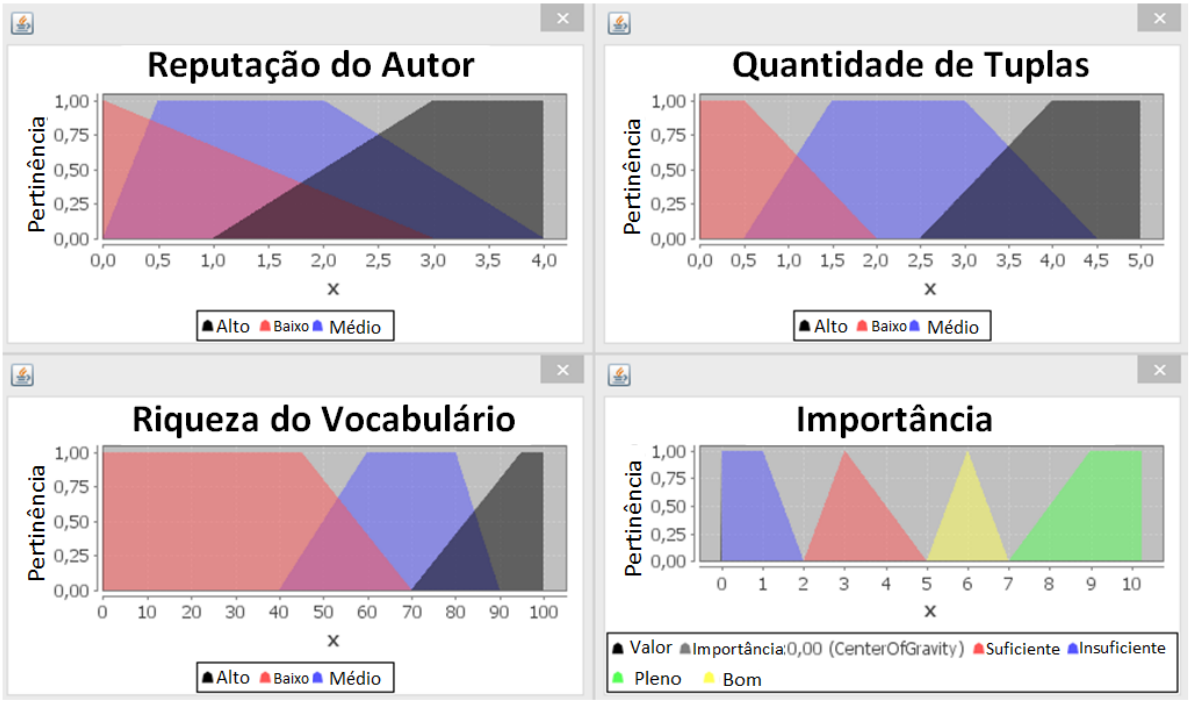

Figura 14 – Funções de pertinência

grande quantidade de valores, e nem muito simplista, pela baixa quantidade de valores. As funções de pertinência para cada valor linguístico podem ser observadas na Figura 14. É importante frisar que os valores das funções de pertinência foram definidos empiricamente pela observação do comportamento das variáveis, nos *corpora* utilizados neste trabalho.

Para a variável de saída (grau de importância), foram utilizados quatro valores linguísticos: **Excelente** (EXC), **Bom** (BM), **Suficiente**(SF) e **Insuficiente** (ISF). Estes valores foram associados a conjuntos *fuzzy* definidos em um universo de discurso U[0,10]. O método de inferência utilizado para a obtenção da saída das regras foi o MAX-MIN e o método de defuzzificação para calcular o valor numérico associado ao grau de importância inferido foi o Centro de Área (*Center of Area*). Esses métodos foram utilizados por terem obtido resultado satisfatório durante as experimentações.

Tabela 6 – Base de regras

	Padrões/Riqueza								
Autor	**B/B**	B/M	B/A	M/B	M/M	M/A	A/B	A/M	A/A
B	**ISF**	ISF	SF	SF	SF	SF	SF	BM	EXC
M	ISF	ISF	SF	SF	SF	BM	BM	EXC	EXC
A	ISF	SF	SF	SF	BM	BM	BM	EXC	EXC

É importante destacar que neste trabalho a dimensão da variável autor não fora explorada com detalhes. Esse fato pode ser comprovado pelas definições da base de regras,

por exemplo, na Tabela 6 todos os elementos da coluna 1 possuem valor **ISF**, independente da qualidade do autor. Isso se repete com as colunas 3 e 4 para o valor **SF** e a última coluna com o valor **EXC**. Certamente, ao considerarmos métricas mais precisas para definir a reputação do autor, os resultados sofrerão alterações.

A Base de Regras *Fuzzy* é um conjunto de regras de produção, que determina a estratégia de tomada de decisão para uma determinada aplicação. A estrutura típica de uma regra *fuzzy* é: **SE** $(x = a)$ **E** $(y = b)$ **E** $(z = c)$ **ENTÃO** $(k = d)$, onde, x, y e z são as variáveis de entrada e k é a variável de saída. Utilizando-se as variáveis propostas pela modelagem da abordagem TOP(X), nessa estrutura, x, y, z e k serão, respectivamente, reputação do autor, número de tuplas, riqueza do vocabulário e grau de importância. Como exemplo, ao observar a Tabela 6, e ao utilizar os valores referentes a intersecção da primeira coluna de *Padrões/Riqueza* com a primeira linha de Autor, as entradas são definidas com os valores: Autor=*baixo(B)*, Padrões=*baixo(B)* e Riqueza=*baixo(B)*, a saída é Importância=*INSUFICIENTE(ISF)* e é referente ao valor contido na interseção das referidas colunas. A base de regras proposta foi modelada de acordo com o *corpus* apresentado na seção 1.3.5.2. As regras estão resumidas na Tabela 6.

3.4 Considerações Finais

Este capítulo apresentou a abordagem TOP(X), usada para inferir a importância de comentários sobre produtos escritos por usuários da Web. O processo de inferência parte do princípio que os comentários podem ser comparados entre si e podem ser classificados de acordo com a sua importância ou utilidade. As informações necessárias para a inferência da importância são extraídas das descrições textuais dos comentários e dos autores envolvidos. Estas informações se resumem em três variáveis, a saber: reputação do autor, quantidade de tuplas $\langle característica, palavra\ opinativa \rangle$ e riqueza de vocabulário.

No capítulo foram apresentadas, também, as motivações para o uso de sistemas *fuzzy* na modelagem do problema, bem como das variáveis de entrada do modelo proposto. Com relação a motivação, os trabalhos mais conhecidos na literatura não apresentam a preocupação em analisar os comentários e decidir se um comentário é importante ou útil, pois o foco está em decidir se a opinião é favorável ou desfavorável em relação ao produto analisado.

As três variáveis de entrada foram escolhidas empiricamente ao observar o conteúdo de opiniões na Web. A reputação do autor é justificada pelo fato de haver pessoas que buscam aumentar a popularidade e aceitação geral de alguns produtos, oferecendo informações sem análise alguma, além de existir a diferença de especialidade entre os consumidores em geral, tornando assim, alguns mais analíticos e confiáveis que outros. A quantidade de tuplas foi escolhida por meio dos trabalhos já encontrados na literatura, eles

apresentam propostas de extração de características e qualidades para nortear a análise de sentimentos dos comentários avaliados. Acredita-se que a maior quantidade dessas tuplas deve levar a uma maior cobertura da avaliação do consumidor em relação às características do produto. Por fim, a riqueza de vocabulário é utilizada para favorecer os comentários que foram escritos com maior atenção com relação ao emprego correto da língua portuguesa.

Na parte final do capítulo, foi definida também a modelagem do sistema *Fuzzy* responsável pela agregação das variáveis de entrada e inferência da importância. As três variáveis definidas compõem a entrada do sistema e a variável grau de importância a saída do mesmo. A apresentação da modelagem incluiu os métodos de fuzzificação e defuzzificação, além da base de regras e funções de pertinências relativas às variáveis de entrada e saída.

Destaca-se que o sistema *fuzzy* pode ser modelado com outras variáveis de entrada e pode ser ajustável ao perfil do usuário. Uma variável interessante é o período em que o cometário foi emitido. Por exemplo, alguém pode estar interessado em comentários sobre uma cidade em um determinado período (por exemplo, a cidade de Caruaru-PE em junho ou a cidade de Gramado-RS em Dezembro). Neste caso, usar comentários emitidos em qualquer período do ano pode perturbar o sistema. Outra possibilidade é valorizar comentários de clientes que possuem o mesmo perfil de quem realiza a consulta. Investigações sobre Sistemas *Fuzzy* Adaptáveis com informações temporais e perfil do usuário estão sendo explorados por um aluno de mestrado do PPGCC/UFPI.

4 Experimentos e Resultados

Este capítulo tem como objetivo apresentar o planejamento, execução e resultados dos procedimentos de experimentação da abordagem proposta por este trabalho.

4.1 Planejamento dos Experimentos

4.1.1 Objetivos

A abordagem TOP(X) é baseada em sistemas *Fuzzy* e tem como objetivo estimar o grau de importância de comentários gerados por usuários. O modelo da abordagem possui três variáveis de entrada: reputação do autor, número de tuplas *<característica, palavra opinativa>* e a riqueza do vocabulário; e uma variável de saída: grau de importância do comentário. Essas três variáveis foram escolhidas para garantir a seleção dos comentários que possuem mais informações para os consumidores. Para garantir o objetivo destacado, conduziu-se um processo de experimentação buscando avaliar a relação de importância dos comentários com sua respectiva quantidade de informações.

Para chegar a um entendimento completo dos objetivos da experimentação, é necessária a explicação de alguns pré-requisitos:

- A abordagem TOP(X) indica a importância de comentários de usuários na Web; na escala [0..10] de números reais.
- Os comentários com a importância inferida são classificados (*ranking*);
- O *ranking* criado pelas importâncias serve para subsidiar decisões de clientes e aos métodos de definição de orientação semântica. Em ambos os casos somente devem ser utilizados os TOP(X) comentários mais importantes.

Dessa forma os objetivos dessa experimentação detalhadamente são:

1. Avaliar se a abordagem TOP(X) seleciona os comentários mais importantes, ou seja, seleciona os comentários mais informativos e úteis aos consumidores;
2. Avaliar se as informações disponibilizadas pelos melhores comentários são suficientes para nortear decisões;
3. Avaliar se a exclusão de comentários menos importantes não afetará o resultado final da análise de orientação semântica;

4. Avaliar se utilizar somente o conjunto de comentários mais importantes pode melhorar os resultados de métodos já existentes.

4.1.2 Fases da Experimentação

A abordagem TOP(X) utiliza um sistema de inferência *fuzzy* para definir a importância dos comentários de consumidores. Como explicado no Capítulo 1, os sistemas de inferência *fuzzy* possuem dois pontos de ajuste: funções de pertinência das variáveis de entrada e saída e base de regras. Portanto, a experimentação foi dividida em duas fases:

1. Fase de ajuste do sistema;
2. Experimentação completa.

Na fase de ajuste, o modelo do sistema *fuzzy* foi exaustivamente modificado. Nessa fase o sistema foi aplicado a um *corpus* criado manualmente, especialmente construído e avaliado para ser utilizado como referência. A Seção 1.3.5.2 descreve em detalhes esses *corpus*. A seção seguinte explica detalhadamente como se deu o processo de ajuste. É importante deixar claro que a referida seção apresenta apenas o resultado final do ajuste.

Na fase de experimentação, a modelagem do sistema já está completa. Dessa forma, a abordagem TOP(X) foi aplicada sobre um *corpus* revisado (ver Seção 1.3.5.1), com o intuito de chegar a conclusões condizentes com os objetivos explanados anteriormente. Durante o experimento, uma das preocupações foi decidir um ponto de corte suficientemente plausível para os objetivos traçados.

4.2 Ajustando o Sistema *Fuzzy*

Como existe a necessidade de ajustar a base de regras e, principalmente, as funções de pertinências das variáveis de entrada e saída, utilizou-se os conhecimentos de um especialista da área de PLN/Ciência da Computação. O processo de ajuste foi realizado sobre o *corpus* manualmente analisado e anotado por um especialista descrito na Seção 1.3.5.2. A distribuição dos comentários do *corpus* pode ser visto novamente na tabela 7.

Tabela 7 – Distribuição da análise manual por níveis de importância

Grau de Importância	Positivos	Negativos
Excelente	8	7
Bom	46	49
Suficiente	80	81
Insuficiente	34	45

O procedimento de ajuste consistiu em extrair os valores das variáveis de cada comentário e executar a abordagem com esses valores. Para cada execução, observa-se o resultado da variável de saída. Em seguida, o sistema foi ajustado exaustivamente a fim de obter melhores resultados. Em cada iteração foram selecionados os comentários mais importantes do *corpus*, nesse caso os comentários *excelentes* e *bons*. Eles ainda foram separados de acordo com a polaridade. Os comentários extraídos foram comparados com os comentários de referência, que são os comentários identificados como *excelentes* e *bons* na análise manual, divididos em quatro conjuntos também de acordo com a polaridade, a saber:

- **Excelentes Positivos:** 8 comentários;
- **Excelentes Negativos:** 7 comentários;
- **Excelentes e Bons Positivos:** 54 comentários;
- **Excelentes e Bons Negativos:** 56 comentários.

Dessa forma, os comentários considerados excelentes pela abordagem foram comparados aos oito comentários *excelentes* positivos determinados pela avaliação manual, ou seja, o conjunto de comentários *excelentes* positivos. Por exemplo, a abordagem considerou 26 comentários positivos como *excelentes*, porém, acertou cinco desses comentários, logo, a precisão para os *excelentes* positivos foi de 19,23% ($5/26 = 0,1923$). Já com relação ao *recall*, o resultado foi 62,5% ($5/8 = 0,6250$). Da mesma forma, para os *excelentes* e *bons* positivos, a abordagem extraiu 60 comentários positivos como *excelentes* e *bons*, porém, desses acertou 36, logo, a precisão para os *excelentes* e *bons* positivos foi de 60% ($36/60 = 0,6$). Com relação ao *recall*, o resultado foi 66,7% ($36/54 = 0,6667$). Esse processo acontece para os outros dois subconjuntos. O intuito dessa comparação é verificar se para o conjunto controlado a abordagem é capaz de encontrar os melhores resultados com os ajustes realizados, a Tabela 8 resume e mostra os valores encontrados. Além das métricas de *precision* e *recall* já citadas, os resultados são apresentados também em relação à métrica *f-measure*.

Tabela 8 – Resultados do procedimento de ajuste

	Positivos				**Negativos**		
	P (%)	**R (%)**	**F (%)**		**P (%)**	**R (%)**	**F (%)**
TOP(8)	19,23	62,5	29,41	**TOP(7)**	0	0	0
TOP(54)	60	66,7	63,17	**TOP(56)**	52,5	37,5	43,75

É possível perceber que na tabela que os melhores resultados foram encontrados com os comentários Excelentes e Bons, sendo que os Positivos apresentaram resultados

de mais de 60% de *F-Measure*. Infelizmente, para os TOP(7) comentários importantes negativos, não foi possível encontrar nenhum elemento. O resultado apresentou valores baixos pois houve uma grande quantidade de características implícitas utilizadas nos comentários dessa classe. Por exemplo, expressões como: "*É melhor migrar para outra marca*", "*congela do nada*", "*câmera mais ou menos*", dentre outras. Essas expressões não permitem a extração efetiva de tuplas por meio dos métodos utilizados, portanto, tornando a extração baixa e assim os comentários não são considerados excelentes. Percebe-se também que quando houve a união dos comentários excelentes com os bons o método foi capaz de encontrar boa parte dos comentários importantes.

4.3 Experimento e Resultado

Após o ajuste do sistema *fuzzy*, fez-se um experimento para avaliar a abordagem. Diferentemente do método de ajustes, a etapa de experimentação foi executada sobre o *corpus* inteiro e revisado, apresentado na Seção 1.3.5.1. É importante relembrar que o *corpus* possui ao todo 1620 comentários, com 982 positivos, 594 negativos e 44 neutros. Na experimentação utiliza-se somente os comentários negativos e positivos devido a pequena quantidade de comentários neutros.

Como foi explicado nos objetivos deste capítulo, a experimentação será feita por meio da comparação entre os resultados dos métodos de detecção da orientação semântica sobre o *corpus* inteiro e um *subcorpus* contendo somente os comentários mais importantes selecionados pela abordagem proposta.

Para inferir a orientação semântica do *corpus*, foi utilizado o léxico de sentimentos *SentiLex-PT* (SILVA; CARVALHO; SARMENTO, 2012). Para extrair os termos a serem polarizados pelo *SentiLex-PT*, foram utilizados os padrões linguísticos explicados na Seção 3.2.2. Sendo assim, para cada comentário foram extraídos os pares $\langle característica, palavra\ opinativa \rangle$, e cada par foi avaliado com a ajuda do *SentiLex-PT*.

De modo a enriquecer o método de inferência semântica, uma lista de advérbios foi coletada do *corpus*. Após isso, foram atribuídos manualmente pesos aos advérbios de acordo com a possível taxa de intensificação ou modificação sobre os qualificadores. Por exemplo, na expressão *A bateria é muito boa*, de acordo com o *SentiLex-PT*, a palavra *boa* possui uma orientação semântica igual a 1 (positivo). À palavra *muito* foi atribuído peso igual a 0.5, portanto, a orientação semântica desta expressão torna-se igual a 1.5. Como outro exemplo, na expressão *A bateria é muito ruim*, a palavra *ruim* possui orientação semântica igual a -1 (negativo). Como o advérbio *muito* possui peso igual a 0.5, então, a orientação semântica da expressão é igual a -1.5. Finalmente, a orientação semântica de um comentário é igual a soma de todos os valores de todos os padrões existentes no

comentário. O processo de definição da orientação semântica das tuplas é especificado no Algoritmo 1. A orientação semântica de um comentário é a soma das orientações da suas tuplas componentes.

Alguns trechos do Algoritmo 1 devem ser explanados para um melhor entendimento. As linhas 2 e 3 definem as listas de qualificadores e modificadores existentes na tupla, pois qualificadores e modificadores não são unitários em todas as tuplas. As polaridades dos qualificadores existentes na tupla serão somados para gerar um valor global de polaridade (linhas 8 - 10). Para os modificadores, todos os pesos dos advérbios existentes na tupla também serão somados (linhas 11 - 13). Para calcular a polaridade efetiva da tupla, são consideradas dois cenários, se a tupla possui ou não pelo menos um modificador, se houver, o inversor será aplicado aos pesos dos modificadores, caso contrário, o inversor será aplicado apenas no total dos qualificadores (linhas 14, 20 - 22). Por fim, a aplicação dos modificadores será determinada pela polaridade dos qualificadores da tupla, se positiva, o total dos modificadores será somado, se negativa, o total dos modificadores será subtraído (linhas 15 - 19), porém, caso haja inversor, esse comportamento será invertido.

Algoritmo 1: Processo de definição da orientação semântica de uma tupla

Entrada: uma tupla
Saída: polaridade da tupla

```
1  início
2      qualificadores = qualificadores existentes na tupla
3      modificadores = modificadores existentes na tupla
4      inversor = 1
5      se tupla contém inversor então
6          inversor = -1
7      fim
8      para cada qualificador q_n em qualificadores faça
9          total_qualificadores += polaridade de q_n no SentiLex-PT
10     fim
11     para cada modificador m_n em modificadores faça
12         total_modificadores += peso de m_n na lista de advérbios
13     fim
14     se total_modificadores > 0 então
15         se total_qualificadores > 0 então
16             retorna total_qualificadores + (inversor*total_modificadores);
17         senão
18             retorna total_qualificadores - (inversor*total_modificadores);
19         fim
20     senão
21         retorna inversor*total_qualificadores;
22     fim
23     retorna 0
24 fim
```

O método de orientação semântica foi executado sobre todo o *corpus* e obteve *Precision* de 71.28% para os comentários positivos, e 49.29% para os negativos. Considerando *F-Measure* os valores foram 58.28% e 30.85%, respectivamente para os positivos e negativos. A Tabela 9 apresenta os resultados em termos de *Precision* (P), *Recall* (R) e *F-Measure* (F) que foram utilizados como referência de comparação para os resultados do experimento. É importante deixar claro que o resultado do método sobre o *corpus* completo é usado como referência para os pequenos cortes aplicados na abordagem. Ou seja, as métricas do subconjunto gerado após a realização de um corte são comparadas com as métricas do *corpus* integral, com o objetivo de verificar se o subconjunto gera melhores resultados do que todo o *corpus*.

Tabela 9 – Resultado do método para o *corpus* completo

Positivos			**Negativos**		
P (%)	**R (%)**	**F (%)**	**P (%)**	**R (%)**	**F (%)**
71.28	49.29	58.28	84.85	18.86	30.85

Após a obtenção das métricas para o *corpus* inteiro, o passo seguinte foi executar a abordagem TOP(X) sobre o *corpus* para inferir a importância de cada comentário independente de ser positivo ou negativo. Em seguida, os comentários foram ordenados de modo decrescente em relação à importância. Essa ordenação permite que, dado um determinado ponto de corte, por exemplo 15, seja gerado um subconjunto dos 15 comentários mais importantes do *corpus*, os TOP(15). No entanto, não há um ponto de corte específico entre os comentários mais importantes e os menos importantes, portanto, deve ser definido empiricamente. Para nortear a escolha de um ponto de corte, definiu-se o intervalo de valores [10,150], sendo que cada valor desse intervalo foi utilizado como ponto de corte. Dessa forma foram gerados 140 subconjuntos do *corpus* completo. As Figuras 15 e 16 mostram as variações de P, R e F para cada ponto de corte.

Para deixar claro o procedimento de cálculo das métricas, um exemplo faz-se necessário. No TOP(20), a distribuição de polaridades dos comentários selecionados de acordo com a avaliação manual do *corpus* é: quatro positivos, 13 negativos e três neutros, totalizando 20 comentários. De acordo com a polarização automática, a distribuição foi a seguinte: 16 positivos e quatro negativos, também totalizando 20 comentários. Para os positivos, houve o acerto dos quatro comentários, ou seja a precisão nesse caso foi de 25% ($4/16 = 0,25$) e o *recall* foi de 100% ($4/4 = 1$), em resumo 16 extraídos, quatro acertos dentre quatro possíveis. Para os negativos, também houve o acerto de quatro comentários, ou seja a precisão nesse caso foi de 100% ($4/4 = 1$) e o *recall* foi de 30,77% ($4/13 = 0,3076$), em resumo quatro extraídos, quatro acertos dentre 13 possíveis.

A Figura 15 apresenta a variação das métricas P, R e F para os 140 subconjuntos

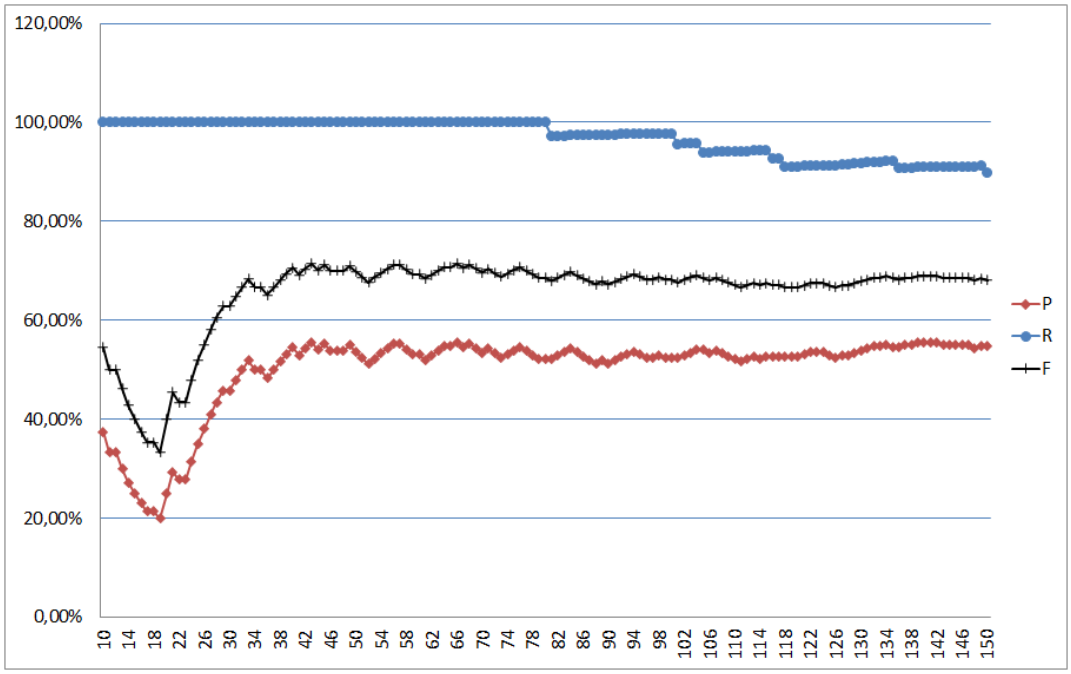

Figura 15 – Variação de P, R e F para os comentários positivos

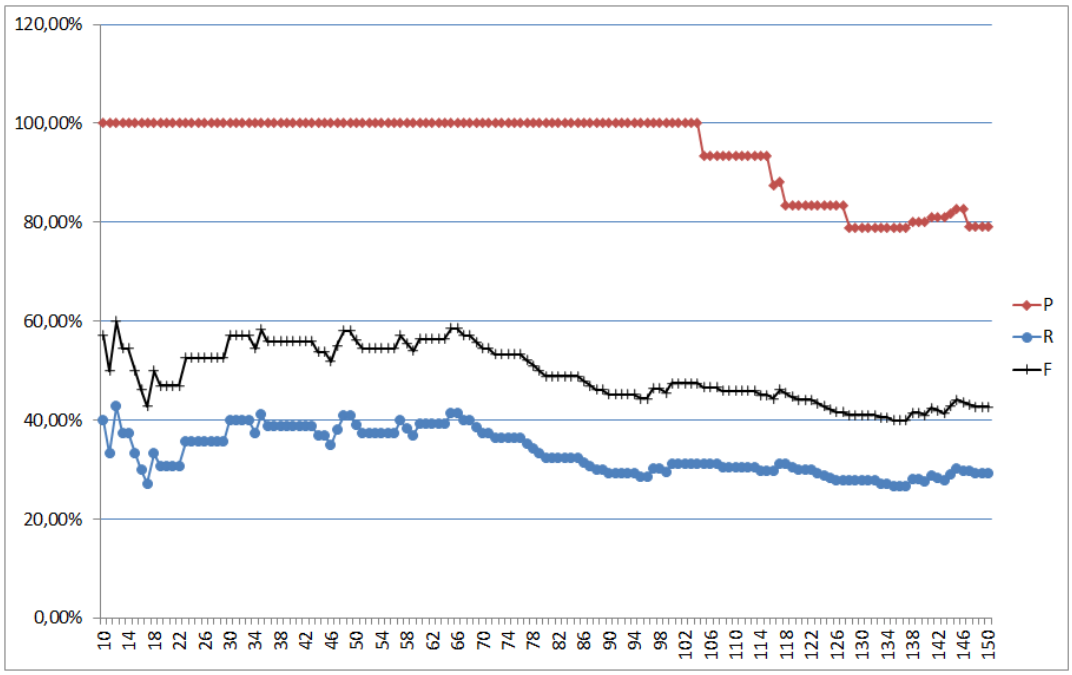

Figura 16 – Variação de P, R e F para os comentários negativos

gerados pelos valores do intervalo definido, observando apenas os comentários positivos. Observa-se que, em termos de precisão, os resultados não foram como o esperado, houve uma diferença de cerca de 15% para o melhor subconjunto em precisão, que foi por volta do TOP(60). Acredita-se que esses valores se deram pela quantidade de comentários considerados positivos pelo método de polarização. Em termos absolutos a quantidade de comentários positivos foi muito maior que a quantidade de comentários negativos, e, portanto, há um percentual maior de comentários positivos no conjunto de comentários

mais importantes. Apesar disso, o *recall* obteve valor de 100% até o TOP(80), esse valor elevado tem relação com a precisão, pois com a maior extração de comentários positivos, o método tem a chance de acertar mais. Em termos de *f-measure*, a partir do TOP(25), a *f-measure* tornou-se maior que o valor medido no *corpus* completo. O maior valor de *f-measure* observado foi exatamente no pico de precisão do TOP(60), aproximadamente 71%, o que representa um ganho de 13% sobre o valor de referência.

Já a Figura 16 é semelhante ao descrito anteriormente, mas considera apenas os comentários negativos. Os resultados para os comentários negativos foram em geral mais satisfatórios que os positivos, pois todas as métricas obtiveram valores mais elevados que a referência. O método obteve 100% de precisão até por volta do TOP(100), significando que a escolha de comentários mais importantes negativos foi efetiva, e, apesar da baixa quantidade de comentários negativos, os selecionados foram os corretamente polarizados. Em termos de *recall* e *f-measure* os resultados foram superiores à referência, apesar do decréscimo por volta do TOP(66). Os valores mais elevados dessas métricas ocorreram também no TOP(66), nesse ponto, os ganhos de *recall* e *f-measure* foram de aproximadamente, 27% e 22%, respectivamente.

Diante dos valores apresentados, observa-se que os melhores pontos de corte estão no intervalo [60,66], onde apesar da perda de precisão para os comentários positivos, houve um acréscimo significativo do restante das métricas para as duas classes, principalmente em relação aos negativos. É importante ressaltar que a tendência dos valores é sempre se aproximar do valor de referência à medida que a quantidade de comentários dos subconjuntos aumenta.

Por fim, destaca-se também alguns problemas que podem ter afetado os resultados dos experimentos:

1. A etiquetagem errada de algumas palavras, pode ter gerado problemas para detectar muitos padrões linguísticos, e por isso, alguns padrões legítimos podem não ter sido extraídos.

2. Apesar de ter sido usado padrões linguísticos adaptados, algumas *tuplas* não foram extraídas corretamente porque elas não possuíam características explícitas ou palavras qualitativas.

3. O valor definido para a variável de entrada “Reputação do Autor” pode ter influenciado os resultados diretamente. A frequência de publicações não é uma métrica perfeita e outros fatores devem ser observados futuramente.

4.4 Considerações Finais

Os principais objetivos e hipóteses que nortearam os experimentos foram apresentados neste capítulo, bem como a descrição do processo de experimentação, que consistiu em duas fases: ajustes e execução da abordagem. A fase de ajustes foi usada para definir os valores finais da base de regras e das pertinências das variáveis do sistema *fuzzy*. Durante os ajustes alguns resultados foram apresentados, indicando que o método não foi muito efetivo sobre comentários negativos, porém, ele atingiu cerca de 60% de *F-Measure* para os comentários positivos.

Após o ajuste, a abordagem foi aplicada ao *corpus* de comentários completo. Descreveu-se ainda, o método de definição de orientação semântica utilizado no experimento. O método foi aplicado sobre o *corpus* completo gerando o resultado das métricas P, R e F, que foram comparadas com os resultados do mesmo método aplicado sobre *subcorpus* de comentários mais importantes. Os *subcorpus* foram definidos por meio de pontos de corte sobre os comentários ordenados de maneira decrescente em relação a importância. Sendo assim, com pontos de corte variando entre 10 e 150, o experimento mostrou que os conjuntos de comentários escolhidos conseguiram obter resultado melhor ou igual à avaliação do *corpus* inteiro, com relação a *f-measure*.

Por fim, foram apresentadas alguns fatores que podem ter influenciado negativamente os resultados do processo de experimentação. Apesar de outros experimentos serem necessários para afirmar categoricamente que a abordagem TOP(X) pode melhorar o processo de análise de sentimentos, os resultados obtidos podem ser considerados satisfatórios.

Conclusões

Este trabalho apresentou a abordagem TOP(X) para estimar o grau de importância de comentários sobre produtos gerados por usuários na Web. A abordagem faz uso de um Sistema de Inferência *Fuzzy* composto por três variáveis de entrada: reputação do autor, número de tuplas e a riqueza do vocabulário e uma variável de saída: grau de importância do comentário. Destaca-se que o diferencial da nossa proposta é aplicar um sistema fuzzy para inferir a importância dos comentários, e definir o conjunto dos TOP(X) mais importantes, minimizando, portanto, o custo da fase de análise de orientação semântica de um produto.

Os comentários mais importantes possuem maior cobertura de opiniões dos consumidores em geral. Portanto, utilizar somente os comentários mais importantes resultará no mesmo desempenho, em termos de acerto, que utilizar todos os comentários possíveis para um determinado domínio.

Discutiu-se também um experimento a fim de avaliar toda a abordagem. Os experimentos realizados mostraram que a abordagem melhorou os resultados do método de definição da orientação semântica, considerando menos de 10% dos comentários do *corpus* original.

Para os experimentos definiu-se um *corpus* de referência para nortear os passos do processo, bem como o ajuste do sistema de inferência *fuzzy*. Assim, o experimento foi conduzido selecionando pontos de corte contendo os TOP(X) comentários mais importantes com X definido no intervalo de 10 a 150.

Os experimentos revelaram que a abordagem proposta melhorou os resultados do método de orientação semântica com o uso de menos de 10% do *corpus* completo. Novos experimentos devem ser realizados, porém os resultados preliminares são encorajadores.

Pela observação dos resultados preliminares algumas afirmações podem ser realizadas:

- Comentários mais importantes selecionados pela abordagem tem maior cobertura de opinião que comentários menos importantes;
- É possível utilizar um pequeno percentual de um conjunto de comentários para retirar as informações mais importantes sobre produtos;
- A utilização de comentários importantes pode ser utilizada para a remoção de ruídos nos métodos de análise de sentimento.

Como contribuições diretas desta Dissertação, destaca-se:

- Abordagem TOP(X) para inferir os melhores comentários de um *corpus* sobre um produto ou serviço, baseada em sistemas *Fuzzy*;
- *Corpus* de *reviews* sobre smartphones do site *buscape.com*, anotados manualmente com relação à polaridade, com cerca de 1620 comentários, sendo 982 positivos, 594 negativos e 44 neutros;
- *Corpus* de *reviews* sobre smartphones do site *buscape.com*, anotados manualmente com relação a importância, com cerca de 350 comentários, sendo 15 excelentes, 95 bons, 161 suficientes e 79 insuficientes.

Como trabalhos futuros, podem ser citados:

- Definir outras métricas para calcular a reputação do autor de forma mais precisa, e analisar o impacto dessas medidas no modelo proposto;
- Aplicar a abordagem em outros domínios (livros, filmes, eletrodomésticos, viagens, hotéis, dentre outros), para analisar o grau de dependência da proposta;
- Realizar uma comparação da abordagem TOP(X), que usa sistemas *fuzzy*, com modelos que utilizam outros métodos de aprendizagem de máquina (ex: Redes Neurais Artificiais) para analisar métricas de desempenho e precisão.
- Investigar o uso de uma variável para tratar questões relativas ao tempo que a opinião foi emitida;
- Definir um Sistema *Fuzzy* Adaptável ao domínio da aplicação e aos interesses específicos do usuário.

Por fim, é importante destacar que este trabalho faz parte de um projeto mais amplo que visa analisar informações de produtos, a partir de três tipos fontes: sites de fabricantes de produtos, sites de vendas e o sites de reclamações. Dessa forma, busca-se gerar um conhecimento mais amplo ao avaliar o produto oferecido por uma empresa, por meio da comparação entre opiniões comuns e opiniões notadamente negativas. Destaca-se também que outros alunos de mestrado do PPGCC/UFPI já estão realizando pesquisas nas dimensões de reputação do autor e sistemas *fuzzy* adaptáveis.

Referências

ACIAR, S. et al. Informed recommender: Basing recommendations on consumer product reviews. *IEEE Intelligent Systems*, IEEE, v. 22, n. 3, p. 39–47, 2007. Citado 2 vezes nas páginas 31 e 42.

AIRES, R. V. X. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil. 2000.* Tese (Mestrado) — Instituto de Ciências Matemáticas, Universidade de São Paulo, São Paulo, 2000. Citado na página 19.

AL-MAIMANI, M.; SALIM, N.; AL-NAAMANY, A. M. Semantic and fuzzy aspects of opinion mining. *Journal of Theoretical and Applied Information Technology*, v. 63, n. 2, 2014. Citado 3 vezes nas páginas 1, 16 e 37.

ALLEN, J. *Natural language understanding*. [S.l.: s.n.], 1995. v. 2. Citado na página 11.

ALMEIDA, P. E. M. d.; EVSUKOFF, A. G. Sistemas fuzzy. *Sistemas Inteligentes: Fundamentos e Aplicações, 1. Ed. São Paulo: Manole*, p. 89–114, 2003. Citado na página 17.

ANTIQUEIRA, L. *Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos*. Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 21.

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: *LREC*. [S.l.: s.n.], 2010. v. 10, p. 2200–2204. Citado na página 3.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. *Modern information retrieval*. [S.l.]: ACM press New York, 1999. v. 463. Citado na página 9.

BARBOSA, A. P. Teoria e prática dos sistemas de classificação bibliográfica. Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 1969. Citado na página 10.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing: O'Reilly, 2009. ISBN 978-0-596-51649-9. Disponível em: <http://www.nltk.org/book>. Citado na página 7.

CINGOLANI, P.; ALCALA-FDEZ, J. jfuzzylogic: a robust and flexible fuzzy-logic inference system language implementation. In: *FUZZ-IEEE*. [S.l.: s.n.], 2012. p. 1–8. Citado 4 vezes nas páginas 15, 22, 23 e 24.

CONRADO, M. da S. et al. A survey of automatic term extraction for brazilian portuguese. *Journal of the Brazilian Computer Society*, Springer, v. 20, n. 1, p. 1–28, 2014. Citado na página 37.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: ACM. *Proceedings of the 12th international conference on World Wide Web*. [S.l.], 2003. p. 519–528. Citado na página 3.

DING, X.; LIU, B.; YU, P. S. A holistic lexicon-based approach to opinion mining. In: ACM. *Proceedings of the 2008 International Conference on Web Search and Data Mining*. [S.l.], 2008. p. 231–240. Citado na página 30.

ESULI, A.; SEBASTIANI, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In: *Proceedings of Language Resources and Evaluation Conference (LREC)*. [S.l.: s.n.], 2006. v. 6, p. 417–422. Citado na página 32.

FELLBAUM, C. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication). illustrated edition edn*. [S.l.]: The MIT Press, 1998. Citado na página 32.

FERREIRA, A. d. H. *Novo dicionário Aurélio da língua portuguesa*. [S.l.]: Editora Positivo, 2004. Citado na página 7.

FREITAS, C. et al. Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. *ENCONTRO DE LINGUÍSTICA DE CORPUS*, v. 11, 2012. Citado na página 3.

FU, G.; WANG, X. Chinese sentence-level sentiment classification based on fuzzy sets. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. [S.l.], 2010. p. 312–319. Citado na página 34.

GALISSON, R.; COSTE, D. *Dicionário de didáctica das línguas*. [S.l.: s.n.], 1983. Citado na página 21.

GAMALLO, P.; GARCIA, M. Freeling e treetagger: um estudo comparativo no âmbito do português. 2013. Citado na página 20.

GHOSE, A.; IPEIROTIS, P. G. Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality. In: *Proceedings of the 16th Annual Workshop on Information Technology and Systems*. [S.l.: s.n.], 2006. p. 1–25. Citado na página 40.

GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. Large-scale sentiment analysis for news and blogs. *International Conference on Web And Social Media - ICWSM'2007*, v. 7, 2007. Citado na página 32.

HEDLEY, J. *jsoup: Java html parser*. 2010. Citado na página 25.

HU, M.; LIU, B. Mining opinion features in customer reviews. In: *Association for the Advancement of Artificial Intelligence (AAAI)*. [S.l.: s.n.], 2004. v. 4, p. 755–760. Citado na página 29.

IEC. International electrotechnical commission technical committee industrial process measurement and control. *IEC 61131 - Programmable Controllers - Part 7: Fuzzy control programming*, 2000. Citado na página 23.

INDURKHYA, N.; DAMERAU, F. J. *Handbook of Natural Language Processing*. 2.. ed. Boca Raton, FL: CRC, 2010. ISBN 978-1-4200-8592-1. Citado 2 vezes nas páginas 8 e 11.

JACKSON, P.; MOULINIER, I. *Natural language processing for online applications: Text retrieval, extraction and categorization*. [S.l.]: John Benjamins Publishing, 2007. v. 5. Citado na página 7.

JEONG, H.; SHIN, D.; CHOI, J. Ferom: Feature extraction and refinement for opinion mining. *Electronics and Telecommunications Research Institute (ETRI) Journal*, Electronics and Telecommunications Research Institute, v. 33, n. 5, p. 720–730, 2011. Citado 2 vezes nas páginas 31 e 32.

JINDAL, N.; LIU, B. Opinion spam and analysis. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2008. (WSDM '08), p. 219–230. ISBN 978-1-59593-927-2. Citado na página 41.

JUSOH, S.; ALFAWAREH, H. M. Applying fuzzy sets for opinion mining. In: IEEE. *International Conference on Computer Applications Technology (ICCAT), 2013*. [S.l.], 2013. p. 1–5. Citado na página 36.

KAMPS, J. et al. Using wordnet to measure semantic orientations of adjectives. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. [S.l.]: European Language Resources Association (ELRA), 2004. Citado 2 vezes nas páginas 32 e 33.

KAO, A.; POTEET, S. R. *Natural language processing and text mining*. [S.l.]: Springer, 2007. Citado na página 8.

KAR, A.; MANDAL, D. P. Finding opinion strength using fuzzy logic on web reviews. *International Journal of Engineering and Industries*, v. 2, n. 1, p. 37–43, 2011. Citado 2 vezes nas páginas 35 e 43.

KHAN, K. et al. Mining opinion from text documents: A survey. In: IEEE. *3rd IEEE International Conference on Digital Ecosystems and Technologies, 2009. DEST'09.* [S.l.], 2009. p. 217–222. Citado na página 1.

LEITE, D. S. et al. Extractive automatic summarization: Does more linguistic knowledge make a difference. *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, p. 17–24, 2007. Citado na página 8.

LI, H. et al. Extracting verb expressions implying negative opinions. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2015. Citado na página 34.

LIMA, L. R. S. de. *Categorização automática de documentos médicos*. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2000. Citado na página 10.

LIU, B. Sentiment analysis and subjectivity. *Handbook of natural language processing*, Chapman & Hall, v. 2, p. 627–666, 2010. Citado 5 vezes nas páginas 1, 2, 12, 13 e 15.

LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012. Citado 2 vezes nas páginas 2 e 13.

LIU, J.; WU, G.; YAO, J. Opinion searching in multi-product reviews. In: IEEE. *The Sixth IEEE International Conference on Computer and Information Technology, 2006. CIT'06.* [S.l.], 2006. p. 25–25. Citado 2 vezes nas páginas 29 e 42.

LUCCA, J. D.; NUNES, M. d. G. V. Lematização versus stemming. 2002. Citado na página 21.

MAMDANI, E. H.; ASSILIAN, S. An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, Elsevier, v. 7, n. 1, p. 1–13, 1975. Citado na página 17.

MANI, I. *Automatic summarization*. [S.l.]: John Benjamins Publishing, 2001. v. 3. Citado na página 8.

MARRO, A. A. et al. Lógica fuzzy: conceitos e aplicações. *Universidade Federal do Rio Grande do Norte*, 2010. Citado na página 17.

MARTINS, C. B. et al. Introdução à sumarização automática. *Relatório Técnico RT-DC*, v. 2, p. 2001, 2001. Citado na página 8.

MASON, O. Qtag-a portable probabilistic tagger. 1997. Citado na página 20.

MEJOVA, Y. Sentiment analysis: An overview. *Computer Science Dept., University of Iowa*, 2009. Citado na página 34.

MENDEL, J. M. Uncertain rule-based fuzzy logic system: introduction and new directions. Prentice–Hall PTR, 2001. Citado na página 17.

NADALI, S.; MURAD, M.; KADIR, R. Sentiment classification of customer reviews based on fuzzy logic. In: IEEE. *International Symposium in Information Technology (ITSim), 2010*. [S.l.], 2010. v. 2, p. 1037–1044. Citado na página 35.

PALTOGLOU, G.; GIACHANOU, A. Opinion retrieval: Searching for opinions in social media. In: *Professional Search in the Modern World*. [S.l.]: Springer, 2014. p. 193–214. Citado na página 44.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, Now Publishers Inc., v. 2, n. 1-2, p. 1–135, 2008. Citado 2 vezes nas páginas 1 e 2.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. [S.l.], 2002. p. 79–86. Citado na página 3.

PARDO, T. A. S. *Sumarização automática: principais conceitos e sistemas para o português brasileiro*. [S.l.]: ICMC-USP, 2008. Citado na página 8.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. d. G. V. Gistsumm: A summarization tool based on a new extractive method. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer, 2003. p. 210–218. Citado na página 8.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. d. G. V. Neuralsumm: Uma abordagem conexionista para a sumarizaçao automática de textos. *Anais do IV Encontro Nacional de Inteligência Artificial*, p. 1–10, 2003. Citado na página 8.

PARR, T. *The Definitive ANTLR Reference: Building Domain-Specific Languages*. First. [S.l.]: Pragmatic Bookshelf, 2007. (Pragmatic Programmers). ISBN 0978739256. Citado 2 vezes nas páginas 15 e 22.

PARR, T. *The Definitive ANTLR 4 Reference*. 2nd. ed. [S.l.]: Pragmatic Bookshelf, 2013. ISBN 1934356999, 9781934356999. Citado na página 22.

PENG, T.-C.; SHIH, C.-C. An unsupervised snippet-based sentiment classification method for chinese unknown phrases without using reference word pairs. In: IEEE. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. [S.l.], 2010. v. 3, p. 243–248. Citado na página 42.

RATNAPARKHI, A. *A Maximum Entropy Model for Part-Of-Speech Tagging*. 1996. Citado na página 20.

ROBERTSON, D. A note on the classical origin of "circumstances" in the medieval confessional. *Studies in Philology*, JSTOR, p. 6–14, 1946. Citado na página 7.

SCHMID, H. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. 1994. Citado na página 20.

SEERAT, B.; AZAM, F. Opinion mining: Issues and challenges (a survey). *International Journal of Computer Applications (0975–8887) Volume 49*, n. 9, p. 42–51, 2012. Citado na página 37.

SILVA, B. C. Dias-da; OLIVEIRA, M. F. de; MORAES, H. R. de. Groundwork for the development of the brazilian portuguese wordnet. In: *Advances in natural language processing*. [S.l.]: Springer, 2002. p. 189–196. Citado na página 32.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer, 2012. p. 218–228. Citado 3 vezes nas páginas 3, 32 e 52.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de análise de sentimento no nível de característica. In: *4nd International Workshop on Web and Text Intelligence (WTI'12), Curitiba*. [S.l.: s.n.], 2012. Citado na página 2.

SIMÃO, A. K. G. Lematização de unidades fraseológicas diacríticas em dicionários bilíngues espanhol/português. *Domínios de Lingu@ gem*, v. 8, n. 2, p. 269–288, 2014. Citado na página 21.

SOUZA, M. et al. Construction of a portuguese opinion lexicon from multiple resources. *STIL*, 2011. Citado na página 3.

SRIVASTAVA, R.; BHATIA, M. Quantifying modified opinion strength: A fuzzy inference system for sentiment analysis. In: IEEE. *International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013*. [S.l.], 2013. p. 1512–1519. Citado na página 36.

SUGENO, M.; KANG, G. Structure identification of fuzzy model. *Fuzzy sets and systems*, Elsevier, v. 28, n. 1, p. 15–33, 1988. Citado na página 17.

TAKAGI, T.; SUGENO, M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, IEEE, n. 1, p. 116–132, 1985. Citado na página 17.

TANG, H.; TAN, S.; CHENG, X. A survey on sentiment detection of reviews. *Expert Systems with Applications*, Elsevier, v. 36, n. 7, p. 10760–10773, 2009. Citado na página 37.

TANSCHEIT, R. Sistemas fuzzy. *Inteligência computacional: aplicadaa administraçao, economia e engenharia em Matlab*, p. 229–264, 2004. Citado na página 17.

TUMITAN, D.; BECKER, K. Tracking sentiment evolution on user-generated content: A case study on the brazilian political scene. In: *Brazilian Symposium on Databases (SBBD)*. [S.l.: s.n.], 2013. p. 6. Citado na página 44.

TURNEY, P. D. Thumbs up or thumbs down? In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Morristown, NJ, USA: Association for Computational Linguistics, 2002. p. 417. Citado 2 vezes nas páginas 2 e 43.

TURNEY, P. D.; LITTMAN, M. L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, v. 21, n. 4, p. 315–346, 2003. Citado na página 33.

VINODHINI, G.; CHANDRASEKARAN, R. Sentiment analysis and opinion mining: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, v. 2, n. 6, 2012. Citado 2 vezes nas páginas 34 e 37.

WIEBE, J.; RILOFF, E. Creating subjective and objective sentence classifiers from unannotated texts. In: *Computational Linguistics and Intelligent Text Processing*. [S.l.]: Springer, 2005. p. 486–497. Citado na página 3.

XU, C. Detecting collusive spammers in online review communities. In: *Proceedings of the Sixth Workshop on Ph.D. Students in Information and Knowledge Management*. New York, NY, USA: ACM, 2013. (PIKM '13), p. 33–40. ISBN 978-1-4503-2422-9. Citado na página 42.

ZADEH, L. Fuzzy Sets*. *Information and control*, Elsevier, v. 8, n. 3, p. 338–353, 1965. Citado na página 16.

ZADEH, L. A. The concept of a linguistic variable and its application to approximate reasoning. *Information sciences*, Elsevier, v. 8, n. 3, p. 199–249, 1975. Citado 2 vezes nas páginas 2 e 16.

ZADEH, L. A. Fuzzy logic and approximate reasoning. *Synthese*, Springer, v. 30, n. 3-4, p. 407–428, 1975. Citado na página 16.

ZANATTA, F.; MIRANDA, F. B. A normatividade em dicionários gerais de língua portuguesa. *A língua portuguesa no mundo. São Paulo: FFLECH-USP*, p. 1–20, 2008. Citado na página 21.

ZHANG, L.; LIU, B. Identifying noun product features that imply opinions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. [S.l.], 2011. p. 575–580. Citado na página 2.

Apêndices

APÊNDICE A – Arquivo FCL

```
FUNCTION_BLOCK sentiment

VAR_INPUT
reputacao_autor : REAL;
padroes : REAL;
    riqueza_vocabulario: REAL;
END_VAR

VAR_OUTPUT
importancia : REAL;
END_VAR

FUZZIFY reputacao_autor
    TERM baixo := (0, 1) (3, 0) ;
    TERM medio := (0, 0) (0.5,1) (2,1) (4,0);
TERM alto := (1, 0) (3, 1);
END_FUZZIFY

FUZZIFY padroes
TERM baixo := (0, 1) (0.5,1) (2, 0) ;
TERM medio := (0.5,0) (1.5,1) (3,1) (4.5,0) ;
    TERM alto :=  (2.5,0) (4,1) (5,1);
END_FUZZIFY

FUZZIFY riqueza_vocabulario
TERM baixo := (0, 1) (45, 1) (70,0) ;
TERM medio := (40,0) (60,1) (80,1) (90,0) ;
    TERM alto :=  (70,0) (95,1) (100,1);
END_FUZZIFY

DEFUZZIFY importancia
    TERM insuficiente := (0,1) (1,1) (2,0);
    TERM suficiente := (2,0) (3,1) (5,0);
    TERM bom := (5,0) (6,1) (7,0);
    TERM excelente := (7,0) (9,1) (10,1);
```

```
METHOD : COG;
DEFAULT := 0;
END_DEFUZZIFY

RULEBLOCK No1
AND : MIN;
ACT : MIN;
ACCU : MAX;

    RULE 1 : IF reputacao_autor IS baixo
    AND padroes IS baixo AND riqueza_vocabulario IS baixo
    THEN importancia IS insuficiente;
    RULE 2 : IF reputacao_autor IS baixo
    AND padroes IS baixo AND riqueza_vocabulario IS medio
    THEN importancia IS insuficiente;
    RULE 3 : IF reputacao_autor IS baixo
    AND padroes IS baixo AND riqueza_vocabulario IS alto
    THEN importancia IS suficiente;
    RULE 4 : IF reputacao_autor IS baixo
    AND padroes IS medio AND riqueza_vocabulario IS baixo
    THEN importancia IS suficiente;
    RULE 5 : IF reputacao_autor IS baixo
    AND padroes IS medio AND riqueza_vocabulario IS medio
    THEN importancia IS suficiente;
    RULE 6 : IF reputacao_autor IS baixo
    AND padroes IS medio AND riqueza_vocabulario IS alto
    THEN importancia IS suficiente;
    RULE 7 : IF reputacao_autor IS baixo
    AND padroes IS alto AND riqueza_vocabulario IS baixo
    THEN importancia IS suficiente;
    RULE 8 : IF reputacao_autor IS baixo
    AND padroes IS alto AND riqueza_vocabulario IS medio
    THEN importancia IS bom;
    RULE 9 : IF reputacao_autor IS baixo
    AND padroes IS alto AND riqueza_vocabulario IS alto
    THEN importancia IS excelente;
    RULE 10 : IF reputacao_autor IS medio
    AND padroes IS baixo AND riqueza_vocabulario IS baixo
```

```
THEN importancia IS insuficiente;
RULE 11 : IF reputacao_autor IS medio
AND padroes IS baixo AND riqueza_vocabulario IS medio
THEN importancia IS insuficiente;
RULE 12 : IF reputacao_autor IS medio
AND padroes IS baixo AND riqueza_vocabulario IS alto
THEN importancia IS suficiente;
RULE 13 : IF reputacao_autor IS medio
AND padroes IS medio AND riqueza_vocabulario IS baixo
THEN importancia IS suficiente;
RULE 14 : IF reputacao_autor IS medio
AND padroes IS medio AND riqueza_vocabulario IS medio
THEN importancia IS suficiente;
RULE 15 : IF reputacao_autor IS medio
AND padroes IS medio AND riqueza_vocabulario IS alto
THEN importancia IS bom;
RULE 16 : IF reputacao_autor IS medio
AND padroes IS alto AND riqueza_vocabulario IS baixo
THEN importancia IS bom;
RULE 17 : IF reputacao_autor IS medio
AND padroes IS alto AND riqueza_vocabulario IS medio
THEN importancia IS excelente;
RULE 18 : IF reputacao_autor IS medio
AND padroes IS alto AND riqueza_vocabulario IS alto
THEN importancia IS excelente;
RULE 19 : IF reputacao_autor IS alto
AND padroes IS baixo AND riqueza_vocabulario IS baixo
THEN importancia IS insuficiente;
RULE 20 : IF reputacao_autor IS alto
AND padroes IS baixo AND riqueza_vocabulario IS medio
THEN importancia IS suficiente;
RULE 21 : IF reputacao_autor IS alto
AND padroes IS baixo AND riqueza_vocabulario IS alto
THEN importancia IS suficiente;
RULE 22 : IF reputacao_autor IS alto
AND padroes IS medio AND riqueza_vocabulario IS baixo
THEN importancia IS suficiente;
RULE 23 : IF reputacao_autor IS alto
AND padroes IS medio AND riqueza_vocabulario IS medio
```

```
    THEN importancia IS bom;
    RULE 24 : IF reputacao_autor IS alto
    AND padroes IS medio AND riqueza_vocabulario IS alto
    THEN importancia IS bom;
    RULE 25 : IF reputacao_autor IS alto
    AND padroes IS alto AND riqueza_vocabulario IS baixo
    THEN importancia IS bom;
    RULE 26 : IF reputacao_autor IS alto
    AND padroes IS alto AND riqueza_vocabulario IS medio
    THEN importancia IS excelente;
    RULE 27 : IF reputacao_autor IS alto
    AND padroes IS alto AND riqueza_vocabulario IS alto
    THEN importancia IS excelente;
END_RULEBLOCK

END_FUNCTION_BLOCK
```

APÊNDICE B – Base de Advérbios

Advérbio	**Peso**	**Advérbio**	**Peso**
rápido	0.5	devagar	0.5
muito	0.5	teimosamente	0.5
longe	0.3	imediatamente	0.3
fluidamente	0.5	corretamente	0.5
razoavelmente	0.4	absurdamente	0.8
infelizmente	0.4	extremamente	0.8
espetacularmente	0.9	jamais	-1.0
ocasionalmente	0.3	perto	0.3
facilmente	0.4	infinitamente	0.8
rapidamente	0.6	incrivelmente	0.7
exatamente	0.3	péssimo	0.6
levemente	0.4	não	-1.0
altamente	0.8	dificilmente	0.4
nunca	0.8	talvez	0.3
maravilhosamente	0.9	pouco	0.3
drasticamente	0.7	estupidamente	0.8
raramente	0.9	diariamente	0.9
prontamente	0.5	demais	0.8
relativamente	0.3	breve	0.5
mal	0.9	enormemente	0.5
pouquíssimo	0.7	substancialmente	0.8
quase	0.4	supostamente	0.7
já	0.3	devagar	0.6
tranquilamente	0.7	ridiculamente	0.9
muitíssimo	0.8	moderadamente	0.3
especialmente	0.5	razoavelmente	0.3
eternamente	0.4	bastante	0.6
bem	0.6	baixo	0.5
muitíssimo	0.8	constantemente	0.5
adequadamente	0.4	intensamente	0.5
perfeitamente	0.7	definitivamente	0.3
meio	0.3	profundamente	0.7
completamente	0.7	plenamente	0.7
totalmente	0.8	lentamente	0.8
absolutamente	0.8	realmente	0.6
simplesmente	0.4	constantemente	0.7
mais	0.6	aparentemente	0.3
precariamente	0.6	sempre	0.8
instantaneamente	0.8	demasiadamente	0.9
extremamente	0.9	abaixo	0.3
pior	0.6	impressionantemente	0.9

APÊNDICE C – Base de Características

Características	
qualidade	produto
som	câmera
memória	interface
design	tela
conexão	antena
caneta	fone de ouvido
processador	teclado
imagem	foto
ligação	sinal
durabilidade	uso
preço	custo
música	vídeo
áudio	display
celular	aparelho
design	desempenho
resolução	tamanho
smartphone	espessura
screen	fotografias
valor	característica
armazenamento	bateria

APÊNDICE D – Gramática Sintagmática

```
grammar GramaticaSintagmatica;

tokens {
    ART,ADJ,SUBS,VERB,PREP,PRON,CONJ,ADV,NUM,FABRICANTE
}


WS  :   ( ' ' | '\t' | '\r' | '\n') {skip();} ;

fragment
LetrasAcentuadas :'á'|'à'|'â'|'ã'|'é'|'ê'|'í'|'ó'|
'ô'|'õ'|'ú'|'ü'|'ç'|'À'|'Á'|'Â'|'Ã'|
'É'|'Ê'|'Í'|'Ó'|'Ô'|'Õ'|'Ú'|'Ü'|'Ç'|'-';

PALAVRA : ('a'..'z'|'A'..'Z'|LetrasAcentuadas)+ ;

INTS  : '0'..'9'+ ('.' '0'..'9'+)? ;

TERMINAL : '.'|'?'|'!'|';';
VIRGULA : ',';

MODELO: (('a'..'z'|'A'..'Z') '0'..'9')+
      ;

NAO : 'não';

periodo :(oracao | . )*?
        ;

oracao : sn_suj? nao_opc? sv sp_opc?
       | sp VIRGULA sn_suj? nao_opc? sv sp_opc?
       ;

nao_opc: NAO
       ;
```

```
sn_opc : sn
       ;

sn : SUBS
   | det sa_opc? SUBS mod?
   ;

sn_suj : ( SUBS (PREP? SUBS)* | PRON )
       | det sa_opc? SUBS (PREP? SUBS)* mod?
       ;

det : pre_det? det_base pos_det?
    ;

pre_det:
       ;

det_base: ART
        | PRON
        | NUM
        ;

pos_det: NUM
       | PRON
       ;

mod : sa?
    | sp
    ;

sp_opc : sp
       ;

sp : PREP sn
   | ADV sp_opc?
   ;

sa_opc : sa?
```

```
       ;

sa : intens? ADJ ((CONJ | VIRGULA)? ADJ)* sp_opc?
   ;

intens : ADV
       ;

sv : VERB complemento
   ;

complemento : sa_opc? sn_opc? sp_opc?
            ;
```