UNIVERSIDADE SAGRADO CORAÇÃO

GUILHERME CANAL MARQUES

SISTEMA DE ANÁLISE DE SENTIMENTOS BASEADA EM ANÁLISE SUPERFICIAL DE TEXTO

BAURU
2014

GUILHERME CANAL MARQUES

SISTEMA DE ANÁLISE DE SENTIMENTOS BASEADA EM ANÁLISE SUPERFICIAL DE TEXTO

Trabalho de Conclusão de Curso apresentado ao centro de Ciências Exatas e Sociais Aplicadas como parte dos requisitos para obtenção do título de bacharel em Ciência da Computação, sob orientação do Prof. Me. Patrick Pedreira Silva.

BAURU
2014

Marques, Guilherme Canal.

M3573s

Sistema de análise de sentimentos baseada em análise superficial de texto / Guilherme Canal Marques. -- 2014.

55f. : il.

Orientador: Prof. Me. Patrick Pedreira Silva.

Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade do Sagrado Coração – Bauru – SP.

1. PLN. 2. Mineração de opinião. 3. Análise de sentimentos. I. Silva, Patrick Pedreira. II. Título.

GUILHERME CANAL MARQUES

SISTEMA DE ANÁLISE DE SENTIMENTOS BASEADA EM ANÁLISE SUPERFICIAL DE TEXTO

Trabalho de Conclusão de Curso apresentado ao centro de Ciências Exatas e Sociais Aplicadas como parte dos requisitos para obtenção do título de bacharel em Ciência da Computação, sob orientação do Prof. Me. Patrick Pedreira Silva.

Banca examinadora:

_____
Prof. Me. Patrick Pedreira Silva
Universidade do Sagrado Coração

_____
Prof. Dr. Elvio Gilberto da Silva
Universidade do Sagrado Coração

_____
Prof. Me. Henrique Pachioni Martins
Universidade do Sagrado Coração

Bauru, 08 de dezembro de 2014.

AGRADECIMENTOS

Agradeço ao meu orientador Patrick Pedreira, que me auxiliou e tornou possível a realização deste projeto.

A Francielle Gonçalves Alves, minha namorada que me apoiou, incentivou e me ajudou diretamente com a realização do trabalho, meu muito obrigado.

Aos meus amigos de sala, que estiveram ao meu lado, contribuindo e tornando tudo mais fácil, agradeço.

E agradeço também a todos que direta ou indiretamente colaboraram com a minha formação, e com a realização do presente trabalho.

RESUMO

Nos dias de hoje tornou-se comum e até previsível a disputa acirrada entre as grandes empresas em qualquer ramo no mercado, sobretudo no que diz respeito à imagem de seus produtos perante o público consumidor. Hoje a internet, principalmente por meio de sites de e-commerce, apresenta-se como uma fonte de opiniões que permite pré-julgar a qualidade de um produto, por meio dos comentários de outros consumidores. Apesar de ser considerada uma ótima fonte de opiniões, devido a enorme quantidade de dados disponíveis, e diante da necessidade da classificação de cada opinião encontrada, a ação humana nesta tarefa, apesar de eficiente, pode acabar sendo muito lenta. Assim, para sanar tal problema, nota-se a necessidade de automatizar o processo de classificação, o que pode ser feito, por exemplo, por meio de técnicas de mineração de opinião (opinion mining), que são derivadas da mineração de dados. Neste contexto, o presente trabalho teve como objetivo desenvolver, por meio de tal técnica, um algoritmo para classificação automática de opiniões. Para atingir esse objetivo um corpus de opiniões foi montado e serviu de base para construção de um sistema, que explora as potencialidades dos padrões de superfície de texto. O método proposto utiliza como base a frequência de palavras (unigramas, bigramas e trigramas), no processo de classificação. Os resultados obtidos mostram que o sistema conseguiu atingir uma precisão de 82% na tarefa de classificação de opiniões, mostrando o potencial do método sugerido.

Palavras-chave: PLN. Mineração de Opinião. Análise de Sentimentos.

ABSTRACT

Nowadays it become usual and even predictable the fierce dispute between large firms in any branch market, especially as regards of their images of products before consumer public. The internet today, mostly by e-commerce sites, is introduced as an opinion source that permits prejudge the quality of a product, by others comments from consumers. Despite it is consider a great opinion source, due a huge amount of available data, and before need for classify each found opinion, the human action in this task, even though efficient, can be really slow. So, to remedy this problem, it is noticed the need to automate the classification process, what can be done, for example, via opinion mining techniques, which is derivative of mining data. In this context, the present study aimed to develop via this technique, an algorithm for automatic opinion classification. To reach this aim, an opinion corpus was constructed and it served from basis to build a system that explore the potential patterns of surface text. The proposed method uses as basis the word frequencies (unigrams, bigrams and trigrams) in classification process. The obtained results show that the system could reach an 82% of precision in opinion classification task, demonstrating the potential of the method suggested.

Key-words: PNL. opinion mining. feelings analysis

LISTA DE ILUSTAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Nos dias de hoje tornou-se comum e até previsível a disputa acirrada entre as grandes empresas em qualquer ramo no mercado. Segundo Modé (2010), “nunca foi tão forte na indústria brasileira a percepção de aumento da concorrência”. Essa situação faz com que tais empresas tenham de tomar medidas para conhecer melhor o mercado, visando criar estratégias para melhorar as vendas. Uma delas é entender melhor a opinião de seu público-alvo já que, conforme citado por Baisch et al (2006), “em um período de profundas transformações e concorrência acirrada, a busca do conhecimento do cliente torna-se extremamente relevante para as empresas que almejam a perpetuação no mercado.”.

Sendo assim, é de vital importância para qualquer empresa saber a opinião do público sobre seus produtos já lançados, uma vez que deste modo, torna-se possível identificar quais pontos devem ser mudados e qual a melhor forma de lidar com as insatisfações.

Identificada a necessidade de saber a opinião do público, outra questão a ser discutida é em qual meio buscar tal expressão por parte dos clientes, já que se deve levar em consideração a veracidade e autenticidade de tais dados para que se possa confiar em qualquer resultado proveniente dessas informações. (NANNI; CAÑETE, 2009). Assim, é natural pensar na internet como fonte, uma vez que se tornou indispensável nos dias de hoje e, com o aumento do seu uso e facilidade que ela proporciona para comunicação entre as pessoas, cresceu também a utilização das redes sociais, que tem como objetivo centralizar as relações interpessoais dentro da internet. Tal fato tornou este tipo de mídia um grande centro de vinculação de opiniões. “Numa sociedade sintonizada com a internet, as redes sociais atuam como peça-chave para fortalecer círculos de amizade, conhecer pessoas de diferentes culturas, trocar experiências e compartilhar ideias.”. (NANNI; CAÑETE, 2009).

No cenário atual, tem-se como principal rede social o “Facebook”, por causa da sua facilidade de vinculação de ideias como ferramenta e, principalmente, por causa de sua grande popularidade. “Desde fevereiro de 2004, a rede já recebeu 201.6 bilhões de conexões. Atualmente, ela conta com mais de 1.2 bilhão de usuários ativos mensais (só no Brasil são 83 milhões)” (OS NÚMEROS..., 2014), o

que a torna a escolha mais acertada para busca dos dados, sendo uma excelente fonte para sistemas de processamento de língua natural como, por exemplo, os sistemas automáticos de análise de opinião.

Apesar de ser considerada uma ótima fonte de opiniões, devido a enorme quantidade de dados disponíveis, e diante da necessidade da classificação de cada opinião encontrada, a ação humana nesta tarefa, apesar de eficiente, pode acabar sendo muito lenta. Assim, para sanar tal problema, nota-se a necessidade de automatizar o processo de classificação, o que pode ser feito, por exemplo, por meio de técnicas de mineração de opinião (opinion mining), que são derivadas da mineração de dados. Segundo Becker e Tumitan (2013):

> A mineração de opinião é definida em como qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual.

Neste contexto, o presente trabalho teve como objetivo desenvolver, por meio de tal técnica, um algoritmo para classificação automática de opiniões. Tal sistema foi baseado em um treinamento por meio de opiniões coletadas no site do “Buscapé”.

Se considerarmos o foco particular do processamento da língua portuguesa, serão relevantes as contribuições deste trabalho, não só para a validação de uma metodologia de análise de sentimentos sobre empresas, como também para as próprias tarefas de processamento de conteúdo textual, sobretudo na web, principalmente porque há muito poucos recursos para o processamento da língua portuguesa, e não existem atualmente muitos sistemas específicos de processamento de documentos nessa língua.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Explorar as potencialidades dos padrões de superfície de texto para o desenvolvimento de um protótipo de sistema de análise de sentimentos, usando um repositório com opiniões sobre empresas (Buscapé[1]) como um corpus, contribuindo para os avanços dos estudos nessa área em língua portuguesa.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar um levantamento bibliográfico dos métodos a serem utilizados;
- Montar um corpus linguístico[2] para a realização do processamento desejado, utilizando o site Buscapé;
- Criar clusters (abordagem bag-of-words[3]) de termos positivos e negativos (unigramas, bigramas e trigramas) para classificação de opiniões sobre algum produto;
- Pontuar tais palavras ou expressões conforme sua frequência em determinado tipo de opinião;
- Implementar o protótipo, que processe os documentos segundo a metodologia proposta;
- Testar e interpretar os resultados obtidos a partir do protótipo em funcionamento.

[1] http://www.buscape.com.br/
[2] Corpus linguístico é um conjunto de textos escritos ou falados numa língua que serve como base de análise
[3] Abordagem na qual cada documento é representado como um vetor de palavras que ocorrem no documento

3 INTELIGÊNCIA ARTIFICIAL

A inteligência artificial pode ser definida como a capacidade de uma máquina de pensar de maneira semelhante a um ser humano, sendo que um sistema inteligente seria aquele que tem a capacidade de raciocinar, planejar, resolver problemas, armazenar conhecimento, comunicar-se através de uma linguagem e aprender. (GONGORRA, 2007).

Porém para Alan Turing (1950), o pensamento é uma atividade interior muito especial, sendo impossível descrever cientificamente. A partir de tal pensamento, Turing chegou à conclusão, que para um computador ser considerado de fato inteligente, ele deveria primeiramente ser submetido a um teste prático, onde o computador participaria de uma espécie de “jogo da imitação”, sendo que um interrogador faria perguntas diretamente para um interrogado que não deve ser visto, podendo ele ser um humano ou computador, e para ser aprovado, o interrogador não deve conseguir distingui-lo de um ser humano. (GONGORRA, 2007; NAVEGA, 2000; PEROTTONI et al., 2001; ZILIO, 2009; VON ZUBEN, 2007).

Fato é que nenhum computador conseguiu passar no teste de Turing até os dias de hoje, apesar de serem inegáveis os grandes avanços na área de Inteligência Artificial, com grande número de ramificações e especificidades. E uma destas ramificações que será abordada posteriormente, é o processamento de linguagem natural, surgido com o objetivo de ajudar a satisfazer tal teste, tem grande relevância também em outros objetos de estudo, como por exemplo, a mineração de dados e opiniões, deste modo sendo de vital importância também no presente trabalho.

3.1 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é o tratamento computacional dos aspectos da linguagem humana que leva em consideração formatos, estruturas e contextos. Segundo Rosa (2011, p.137) “[...] o Processamento de Línguas Naturais pode ser definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no dia a dia.”, assim basicamente, pode-se dizer que o PLN visa fazer o computador se comunicar em linguagem humana, não necessariamente em todos os níveis de entendimento, sendo eles fonético ou fonológico (sons das palavras), morfológico (origem de cada palavra), sintático

(papel estrutural das palavras ou sentenças dentro do texto), semântico (significado das palavras e expressões dentro do contexto) e pragmático (mudança do significado das palavras e sentenças em diferentes contextos) (SILVA et al., 2007; DA COSTA; RALHA; RALHA, 2006; ROSA, 2011, grifo nosso).

3.1.1 Histórico

Desde o início da introdução dos computadores ao cotidiano das pessoas, o computador tem passado grandes processos de evolução principalmente quando se trata da interação entre homem e máquina. (SILVA et al., 2007).

O conceito de linguagem de programação surgiu a partir da necessidade de fazer com que as máquinas fossem capazes de entender e executar tarefas, tendo com o passar do tempo um grande desenvolvimento imposto pela necessidade de fugir da lógica computacional e se aproximar da forma humana de compreensão. (SILVA et al., 2007).

Necessidade essa que também cresceu, levando cada vez mais pesquisadores a estudar a fundo, com o objetivo de criar computadores cada vez mais inteligentes que pudessem compreender e até mesmo responder à linguagem humana. E assim, hoje esta ramificação do desenvolvimento computacional tornou-se de grande importância, com estudos e pesquisas nas mais diversas áreas do Processamento de Linguagem Natural. (SILVA et al., 2007).

Como citado anteriormente, houve com o passar dos anos, muitas pesquisas e desenvolvimentos nos mais diversos nichos do processamento de linguagem natural, o que acaba por dificultar a ação de determinar um eixo principal sobre a evolução da área. (SILVA et al., 2007).

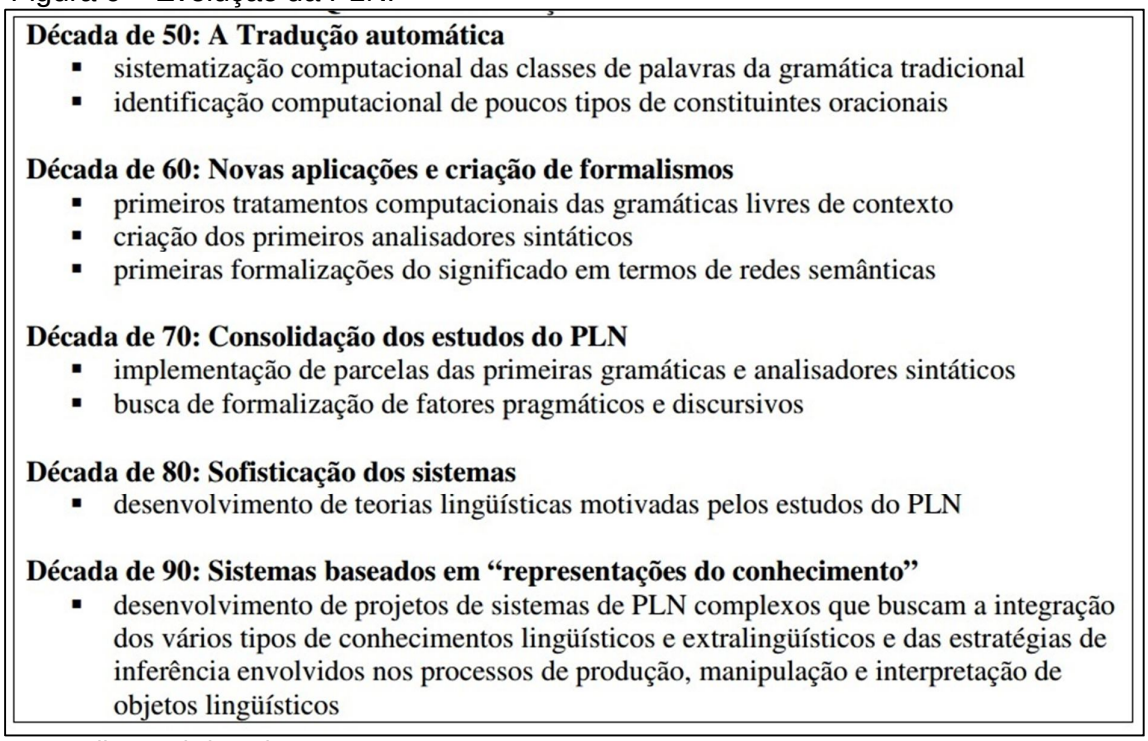
Porém dentre todas as áreas a se levar em consideração, uma se destaca, a tradução automática, uma vez que é considerada pela maioria como o marco inicial na utilização dos computadores para o estudo das línguas naturais, e também pode servir como base para um resumo da evolução de todo o campo. (SILVA et al., 2007).

Os estudos institucionais sobre o PLN tiveram início na década de 1950 nos Estados Unidos, após grande motivação da fundação “Rockfeller” exatamente tratando da tradução automática, o que levou universidades renomadas como MIT e Harvard, a também pesquisar e debater sobre a área. Porém após mais de uma

década de estudos, os resultados continuavam insatisfatórios, por não haver um embasamento linguístico aprofundado, não havia ainda uma tradução automática de fato, os resultados obtidos eram possibilidades de tradução, que posteriormente deveriam ser avaliadas por tradutores humanos, levando a um grande desaquecimento e descredito a área. (SILVA et al., 2007).

Esta realidade foi modificada somente em meados da década de 1970, quando pesquisadores passaram a ter uma atitude mais acadêmica e realista, levando assim, a um número significativo de experiências bem sucedidas, reaquecendo as pesquisas e projetos não apenas sobre o assunto em específico, mas também em outros ramos do processamento de linguagem a partir da década de 1980, tornando um importante objeto de estudo. A Figura 1 mostra um breve resumo dos pontos importantes da evolução da PLN em cada década. (SILVA et al., 2007).

Figura 6 – Evolução da PLN.



Fonte: Silva et al. (2007).
Nota: Adaptada pelo autor.

3.1.2 Aplicações

O Processamento de Linguagem Natural como abordado anteriormente, pode ser ramificado de forma bem ampla e diversificada, apesar ter sua principal origem na aplicação da tradução automática, tem uma grande variedade de aplicações possíveis e interessantes de serem abordadas. (SILVA et al., 2007).

Uma delas é a manipulação de base de dados por meio de uma interface, onde o sistema de processamento de linguagem natural ou simplesmente SPLN tem a função de servir de mediador entre o usuário e a base de dados. Neste tipo de aplicação o sistema irá “traduzir” frases recebidas em linguagem natural, para a linguagem do sistema de gerenciamento dos dados que irá manipular tais informações, num denominado “sistema de perguntas e respostas”. (SILVA et al., 2007; LUGER, 2004, grifo nosso).

Outra aplicação seria em sistemas tutores de forma em que, diferentemente dos sistemas tradicionais de aprendizado por computador, onde as informações são previamente estruturadas e ramificadas pelo projetista do sistema, cria uma interação com o aluno por meio de um sistema inteligente, que usa uma rede de conhecimento composta de fatos, regras e relações entre conteúdos, criando um diálogo com o usuário e, assim, consegue simular a relação entre aluno e professor. (SILVA et al., 2007).

Já em um sistema de automação de tarefas administrativas, o SPLN pode auxiliar em tarefas de rotina e administrativas de uma empresa, dependendo das necessidades internas, pode manipular objetos e ícones do monitor do computador por meio de comandos orais, ou até mesmo responder dúvidas sobre a empresa, ou certa rotina administrativa, entre outras aplicações. (SILVA et al., 2007).

Há também outras inúmeras aplicações possíveis do PLN em um sistema, como em sistemas acadêmicos ou em programação automática, ou até mesmo num sistema de processamento de textos científicos, dando uma amostra de como pode ser utilizado um SPLN. Outra aplicação bastante atual que merece destaque é a utilização de tais sistemas em tarefas de análise de sentimentos, o que justifica a escolha do tema desta investigação.

3.1.3 Construção de um SPLN

Conhecendo as aplicações possíveis e mais relevantes para um sistema de processamento de linguagem natural, é possível entender que ele se trata também de um sistema automático de conhecimentos, uma vez que pode fazer revisões ortográficas, análises sintáticas, fazer perguntas e respostas, entre outras coisas. Deste modo, pode-se entender o estudo do PLN como uma “engenharia do conhecimento linguístico” e, deste modo, utilizar conceitos deste campo para o próprio desenvolvimento. (SILVA et al., 2007).

Assim, pode-se dividir o ato da concepção de um SPLN em três fases, sendo elas a “Fase Linguística”, onde há um estudo mais aprofundado sobre a linguagem, montando-se um corpo de conhecimentos sobre ela, compreendendo todos os fenômenos linguísticos relevantes para o sistema. (SILVA et al., 2007, grifo nosso).

Após isso, há a “Fase Representacional”, que é fase da construção conceitual do sistema, propondo sistemas formais de representação linguística e extralinguística, computacionalmente tratáveis. (SILVA et al., 2007, grifo nosso).

E, por último, a “Fase Implementacional” que codifica todas as representações concebidas na fase anterior em linguagens de programação, além de fazer o planejamento global do sistema. (SILVA et al., 2007, grifo nosso).

Tais fases visam principalmente, dentro da área de estudo do PLN, criar programas que facilitem a comunicação entre o usuário e o computador, utilizando-se de um conjunto de programas capazes de interpretar e gerar informações em mensagens linguisticamente construídas. (SILVA et al., 2007). Um dos recursos mais utilizados para a construção de SPLN são os corpora linguísticos que serão descritos na seção seguinte.

3.1.3.1 Corpus

Corpus é uma palavra originária do latim que significa conjunto, corpo. Sendo usualmente compreendido com o um conjunto temático de dados, informações textuais e documentos. (MELLO; SÁ, 2006; SARDINHA, 2004).

A partir da compreensão e uso de tal definição surge o conceito de corpus linguístico que se baseia na coleta e exploração dos corpora, ou seja, conjuntos de dados linguísticos textuais que são coletados criteriosamente com o propósito de servir para a pesquisa de uma língua ou variedade linguística e, assim, dedicando-se a exploração da linguagem por meio de evidências empíricas, extraídas pelo computador. Pode-se entender então que um corpus linguístico é a representação de uma determinada realidade em determinado tempo, ou seja, é a representação de um contexto previamente definido para a pesquisa. (MELLO; SÁ, 2006; SARDINHA,2004).

3.1.3.2 Abordagem Bag-Of-Words

A abordagem Bag-of-Words surgiu a partir da necessidade de organizar textos obtidos através da mineração de textos. A técnica visa facilitar o trabalho da aprendizagem de máquina, transformando através de um pré-processamento tais textos desestruturados em informações organizadas que poderão ser utilizadas pela maioria dos algoritmos com tal fim. (MARTINS; MATSUBARA; MONARD, 2003).

Tecnicamente, a abordagem Bag-Of-Words receberá o corpus formado na mineração de textos, e atribuirá cada palavra deste texto a um grande vetor, sendo que assim, cada palavra ocupará uma posição deste, facilitando assim a ação computacional em relação à manipulação das palavras, tornando possível então uma classificação de cada grupo de palavras, atribuição de valores pertinentes ao algoritmo que o utilizará, além de uma possível representação visual de tal texto. (MARTINS; MATSUBARA; MONARD, 2003).

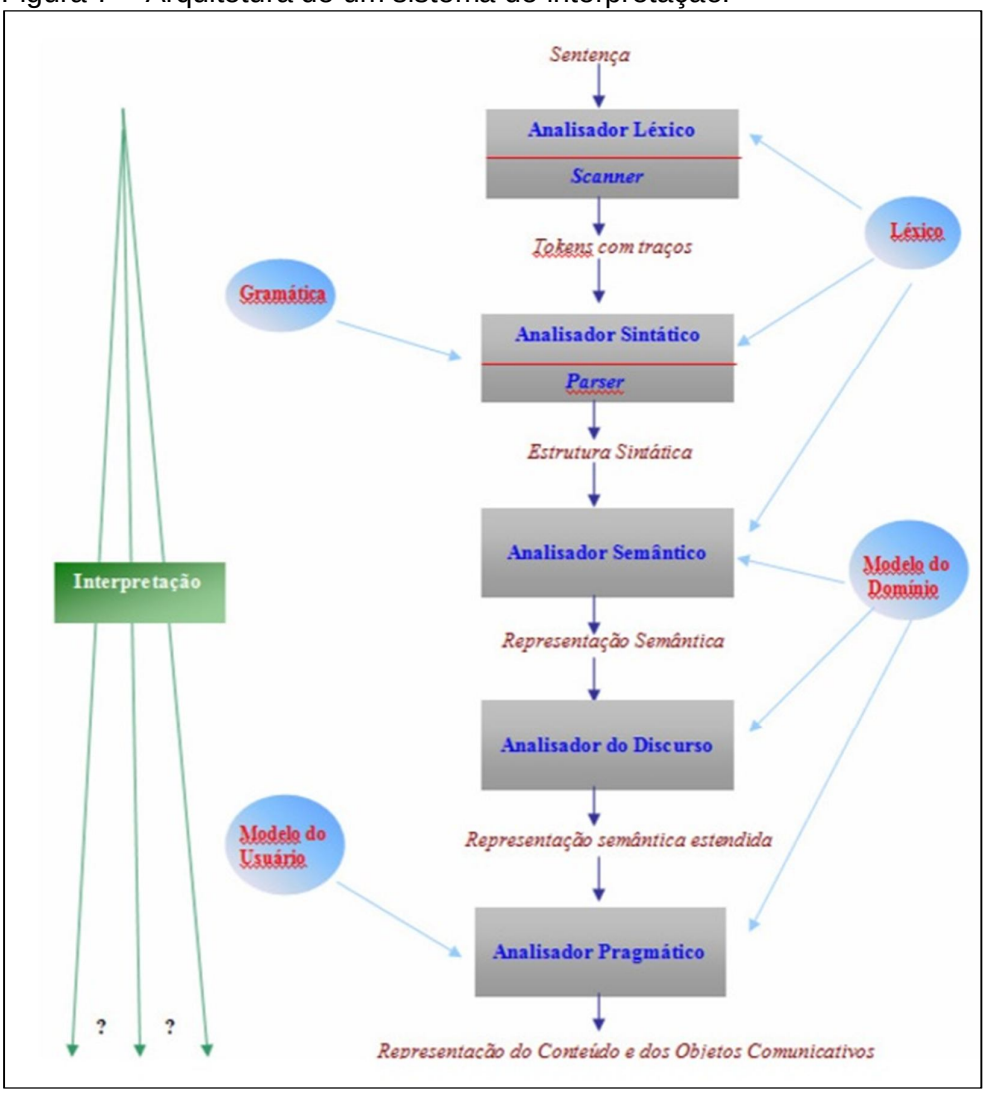
3.1.4 Arquitetura de um SPLN

A arquitetura de um SPLN pode variar de acordo com cada sistema, uma vez que cada aplicação pode demandar um número variado de funcionalidades. Porém, há pelo menos dois elementos básicos a quase todos eles, podendo ocorrer em conjunto ou isoladamente, que são as fases de interpretação e geração, sendo que a primeira faz a transformação da linguagem humana para uma forma mais próxima à linguagem computacional, e a segunda gera sentenças em linguagem humana a partir de dados na linguagem do computador. (SILVA et al., 2007).

3.1.4.1 Arquitetura de um sistema de interpretação

Um sistema de interpretação de língua natural tem sua arquitetura baseada em dois tipos de elementos para ser viável, sendo eles os módulos de processamento, e os recursos e conhecimentos específicos necessários para o processamento, uma vez que cada fase de processamento se baseia em um recurso de conhecimento, como pode ser visto na Figura 2. (SILVA et al., 2007; ROSA, 2011).

Figura 7 – Arquitetura de um sistema de interpretação.



Fonte: Silva et al. (2007).

Como se pode observar através da Figura 2, um sistema de interpretação passa basicamente por cinco processos de análise até chegar ao objetivo, apesar de em alguns nos quais não há a necessidade da interpretação de uma sentença, a arquitetura pode ser simplificada, assim como outros sistemas podem ter a necessidade de um processamento adicional, como uma análise morfológica por exemplo. (SILVA et al., 2007).

Contudo, analisando a Figura 2 é possível notar uma arquitetura simples, que é constituída por processos mais complexos, uma vez que cada tipo de análise leva em consideração vários aspectos para ocorrer. (SILVA et al., 2007).

Assim, seguindo a arquitetura apresentada, a sentença deve primeiramente passar pelo analisador léxico também chamado de scanner, que irá identificar e separar todos os componentes significativos (Tokens), sejam palavras ou símbolos de pontuação, além de associar atributos gramaticais ou semânticos a cada token definindo a relação entre eles, formando uma estrutura linguística. (SILVA et al., 2007; LUGER, 2004, grifo nosso).

Após este processo, a sentença é submetida ao analisador sintático ou parser do latim pars orations (parte do discurso), que deverá recuperar a estrutura sintática, sendo guiado por uma representação gramática da língua, que por sua vez, converte as sentenças em árvores gramaticais, visando os aspectos relevantes à aplicação, para assim diminuir a complexidade do analisador. (SILVA et al., 2007; LUGER, 2004, grifo nosso).

O analisador semântico vem em seguida, sendo responsável pela interpretação da sentença, ou de partes dela, segundo Coppin (2012) “[...] análise semântica é a análise que usamos para extrair significado de uma declaração.”, baseando-se diretamente no Modelo do Domínio, que servirá de consulta, destruindo ambiguidades de sentido, através da combinação entre todos os componentes. (SILVA et al., 2007; COPPIN, 2012).

Havendo a análise semântica, entra em ação o analisador do discurso, que terá o trabalho de analisar o significado de cada sentença como um todo, a partir das sentenças anteriores a ela, criando uma representação expandida do significado das sentenças. (SILVA et al., 2007; COPPIN, 2012).

E por fim, há o analisador pragmático, que analisará as possíveis ambiguidades de sentido das sentenças, uma vez que uma só frase pode ter vários sentidos diferentes (ambiguidade semântica) se analisada de forma isolada, porém havendo sentido claro quando analisada dentro de um contexto mais amplo, o analisador pragmático terá o papel de definir o real sentido de uma sentença ou prover a desambiguação, baseado nos sentidos das sentenças anteriores, ou seja, no contexto, chamado de modelo de mundo . (SILVA et al., 2007; COPPIN, 2012).

3.2 MINERAÇÃO DE DADOS

A mineração de dados (data mining) teve seu início a partir da década de oitenta, quando profissionais perceberam o grande crescimento de dados armazenados nos computadores das empresas, e que não tinham utilidade para elas. Assim, sendo necessária uma forma automática de avaliação de todos os dados pertinentes, para poder classificá-los e, consequentemente, descartar o que não fosse necessário armazenar, economizando espaço de armazenamento, e consequentemente, dinheiro. (AFONSO; GUEDES; MAGALHÃES, 2010).

Posteriormente, com o crescimento da web e das informações que circulam nela, cresceu também a quantidade de dados pesquisados de maneira remota neste meio, e como tais dos estão situados em todas as partes do mundo, notou-se que para viabilizar tal crescimento, como seria de grande utilidade de tal conceito aplicado a este tipo de ambiente, surgindo assim o conceito de web mining. (AFONSO; GUEDES; MAGALHÃES, 2010).

A web mining, tem sua aplicação baseada em três subáreas, a mineração de estruturas, que visa categorizar as páginas web a partir da estrutura de seu link; a mineração de logs, que estuda o comportamento dos usuários em determinado site; e a mineração de conteúdo, que pesquisa os conteúdos de tais páginas, tais como textos, imagens além de outros componentes. (AFONSO; GUEDES; MAGALHÃES, 2010).

É exatamente da última subárea citada que se deriva a mineração de opinião, que visando o texto de determinadas páginas web, servirá como base para o presente trabalho.

3.2.1 Mineração de Opinião

Como dito no tópico anterior, a mineração de opinião deriva da mineração de conteúdo da web mining, que por sua vez surgiu a partir do conceito de mineração de dados, só que aplicada ao ambiente web. (AFONSO; GUEDES; MAGALHÃES, 2010).

Assim como a web mining se diferencia da mineração de dados de acordo com a sua aplicação, a mineração de opinião se diferencia da área de origem, pois através do Processamento de Linguagem Natural e da análise de sentimentos, tem

como objetivo a extração de opiniões e sentimentos de usuários sobre determinados temas ou produtos, contidas em sites especializados, para posteriormente ocorrer a classificação de cada opinião, e descobrir o sentimento associado a cada uma delas. (AFONSO; GUEDES; MAGALHÃES, 2010; BARROS; LIMA; SILVA, 2012).

3.2.1.1 Recuperação da informação

A recuperação da informação é parte vital da mineração de opinião, uma vez que é ela que busca, filtra, armazena e organiza os dados que viabilizam a montagem do corpus linguístico, que serve de base a praticamente todos os sistemas deste seguimento. (AFONSO; GUEDES; MAGALHÃES, 2010).

Deste modo, a recuperação da informação é a área da computação que lida com o armazenamento de documentos e recuperação automática de informações. Basicamente, a recuperação da informação irá buscar os dados necessários a determinado fim em bases de dados, sejam relacionais ou isoladas, e principalmente dentro da web, uma vez que é a maior fonte de dados existente. Deste modo, visando em um maior aproveitamento de resultados, foca em mecanismos de buscas e diretórios de pesquisa, por serem essenciais para mecanismos de recuperação de informação na web. (AFONSO; GUEDES; MAGALHÃES, 2010).

3.2.1.2 Análise de Sentimentos

A Análise de Sentimentos surge praticamente como sinônimo de mineração de opiniões, e tem como objetivo identificar o sentimento expresso em textos opinativos, textos esses que são subjetivos, uma vez que não apresenta fatos concretos como os objetivos. (BARROS; LIMA; SILVA, 2012).

Para chegar a tal objetivo, a Análise de Sentimentos visa indicar a polaridade das opiniões, classificando-as em positivas, negativas ou neutras. Basicamente, a polaridade de um texto é expressa a partir de palavras opinativas, sejam adjetivos, advérbios e até mesmo substantivos. (BARROS; LIMA; SILVA, 2012).

A análise pode ser realizada em três níveis que diferenciam profundidades dentro do texto. O nível de documento, onde observa o sentimento expresso no texto como um todo. Em seguida o nível de sentença classifica a polaridade de cada sentença dentro do texto. E em nível mais profundo, o nível de característica que

indica a polaridade de cada atributo do objeto analisado, conseguindo uma visão refinada de cada opinião. (BARROS; LIMA; SILVA, 2012).

Para conseguir seus resultados, a Análise de Sentimentos se baseia em um processo complexo, sendo que um sistema completo pode ser dividido em quatro etapas. A primeira é a detecção de subjetividade, que irá identificar se um texto é mesmo subjetivo ou objetivo. Em segundo há a extração de características, que identificará as características do objeto analisado, sendo indispensável quando a análise é feita em nível de característica. Para definir a polaridade das opiniões, entra em a ação a etapa de classificação de sentimentos. E a quarta e última etapa é a visualização de resultados, que irá apresentar ao usuário os resultados da análise, seja em gráficos, tabelas, ou até mesmo em linguagem natural. (BARROS; LIMA; SILVA, 2012).

Como é possível notar, é nas duas etapas intermediárias que se dá de fato a análise de sentimentos, tais etapas serão detalhadas nos tópicos a seguir.

3.2.1.2.1 Extração de características

Assim como abordado no tópico anterior, a extração de características tem seu foco nas características do objeto sob análise. Na maior parte das vezes, a extração de características se baseia em comentários online sobre o produto ou serviço em questão. (BARROS; LIMA; SILVA, 2012).

As técnicas empregadas na extração de características variam de acordo com o formato dos comentários que servirão como base. Para os comentários em formato de prós e contras, são necessárias técnicas de aprendizado de máquina, empregada a um corpus com as opiniões já classificadas, conforme a divisão encontrada na fonte. (BARROS; LIMA; SILVA, 2012).

Já para os comentários em texto livre, que contem opiniões positivas misturadas às negativas em sentenças mais elaboradas, existe a utilização de técnicas de linguística e estatística, onde os substantivos mais frequentes no texto são tomados como as características. Posteriormente as palavras opinativas com sentido mais próximo aos dessas características são considerados para características infrequentes. Este método requer um corpus de tamanho considerável, uma vez que para existir uma diferenciação clara entre as características mais frequentes e as menos, é necessário um número grande de

informações. Porém, diferentemente das opiniões de prós e contras, não é necessário haver uma divisão dentro deste corpus. (BARROS; LIMA; SILVA, 2012).

3.2.1.2.2 *Classificação de sentimentos*

A classificação de sentimentos é a principal etapa dentro de uma análise de sentimentos, pois será ela que identificará de fato a polaridade do texto. (BARROS; LIMA; SILVA, 2012).

De acordo com os autores supracitados, assim como na extração de características, a abordagem de aprendizado de máquina também aparece na classificação de sentimentos, necessitando de um corpus de treinamento precisamente dividido e etiquetado, sendo necessário categorizar a polaridade de expressões de menor frequência, o que a torna ao mesmo tempo mais precisa e também muito mais custosa.

Deste modo surge como alternativa, a utilização dos métodos estatísticos e linguísticos, que em geral se baseiam em listas de palavras previamente classificadas, e na falta de tal lista, a tarefa de classificação fica a cargo do usuário, que nem sempre é capaz de realizá-la. Entretanto, existem ferramentas que auxiliam neste tipo de classificação automática. (BARROS; LIMA; SILVA, 2012).

Todo esse processo de análise de sentimentos pode ser realizado no contexto de web mining, permitindo a utilização de sites online para identificação e monitoramento de polaridade em mensagens compartilhadas que carregam emoções expressas pelos usuários. Assim, considerando o foco deste trabalho e o uso potencial do processo de mineração de opinião, é descrita na seção seguinte o site Buscapé, que pode ser considerado uma fonte de opiniões para construção de um corpus voltado ao processo de desenvolvimento de uma ferramenta de análise de sentimentos.

3.2.1.3 Buscapé

O Buscapé é uma empresa brasileira criada em meados de 1999 junto com a popularização da internet no Brasil que tem como principal objetivo proporcionar ao usuário um meio de comparações de preços e serviços. (ARRUDA; PENIDO; ROSSI, 2011).

Criado por três alunos de engenharia de computação da Escola Politécnica da Universidade de São Paulo, o Buscapé tem seu funcionamento baseado na tecnologia Spider, também desenvolvida por eles, que permite um grande armazenamento de informações de sites de e-commerce e as organiza para permitir a comparação. (ARRUDA; PENIDO; ROSSI, 2011).

Seu modelo de negócios é administrado pela “Central de negócios” criada para viabilizar o cadastro de produtos, de forma em que as empresas se associam ao site, para disponibilizarem suas informações, como preço, localidade da loja, formas de pagamento, etc. A empresa apenas paga ao site quando um de seus anúncios é clicado, assim definindo um preço a pagar ao site por cada clique (CPD) numa espécie de “leilão de cliques”, sendo que há um preço mínimo, e quanto mais se investir nele, melhor será a posição da empresa nos resultados de pesquisa. (ARRUDA; PENIDO; ROSSI, 2011).

Hoje em dia o Buscapé já conta com mais de 500 mil empresas listadas no site, além de uma gama de 14 milhões de produtos e número médio de 20 milhões de usuários por mês. Sendo que cada empresa pode ser classificada pelos usuários segundo sua confiabilidade e qualidade de serviço, levando-se em consideração principalmente o prazo de espera pelo produto, e se o usuário voltaria a comprar. Além da opinião, que pode ser deixada abertamente em relação a cada produto listado, as pessoas podem julgar também o serviço da empresa ou um item em específico, seja por sua qualidade em relação aos outros, ou as impressões que obtiveram ao adquiri-lo. (ARRUDA; PENIDO; ROSSI, 2011).

Com milhares de opiniões disponíveis, o Buscapé torna-se um potencial repositório de informações que podem ser exploradas por aplicações de mineração de opinião, o que o torna um excelente objeto de estudo e foco de pesquisas na área de mineração de textos. Essa grande quantidade de opiniões torna impraticável a assimilação por uma pessoa de todas as informações sobre um produto específico, tornando imprescindível uma ferramenta que seja capaz de realizar a

“leitura” desses repositórios ajudando no processo de decisão de compra de um produto. Assim, as opiniões dos usuários deixadas no site podem ser processadas automaticamente, informando ao usuário sobre um conjunto de opiniões que podem estar falando positivamente ou negativamente sobre o produto desejado.

3.3 ENGENHARIA DE SOFTWARE

Segundo Pressman (2011, p.38) “Software tornou-se profundamente incorporado em praticamente todos os aspectos de nossas vidas [...]”, sendo inegável sua ação no cotidiano do homem moderno, porém para que o software seja tão importante, outro conceito teve de surgir. Juntamente com a evolução da utilização do software, que passou a ganhar mais importância a partir dos anos 60 com o surgimento dos sistemas operacionais com características de multiprogramação, e com o crescimento da utilidade e eficiência da computação, o conceito de Engenharia de Software nasce a partir da necessidade de produção de softwares mais complexos em relação às aplicações simplistas que vinham sendo desenvolvidas até então. (MAZZOLA, 199-?).

“Software, em todas suas formas e em todos os seus campos de aplicação, deve passar pelos processos de engenharia.” afirma Pressman (2011, p. 39). Assim a Engenharia de Software surge para mostrar que um programa é mais do que um simples “conjunto de instruções que executadas juntamente, produzem uma determinada função”, para ser tratado então como um produto. Para Sommerville (2007, p.12) “Os produtos de software consistem em programas desenvolvidos e documentação associada.”, uma vez que este produto não será destinado ao programador, e sim ao grande público, que no geral tem pouco conhecimento desta área, gerando também grande preocupação com a interface, além de uma grande bateria de testes para identificar e corrigir todos os possíveis erros, uma vez que um usuário comum não seria capaz de tal tarefa, garantindo assim ao usuário um desfruto máximo do produto adquirido. (MAZZOLA, 199-?).

Deste modo, pode-se entender que a Engenharia de Software tem como objetivo prover a tecnologia necessária para produzir um software de alta qualidade e com baixo custo, a partir de um conjunto de atividades em ordem específica, visando o produto final desejado. Existem três principais atividades comuns a todos

os modelos de processos de desenvolvimento: a fase de definição; a fase de desenvolvimento; e a fase de manutenção. (MAZZOLA, 199-?).

Na Fase de Definição será definido o que será feito no projeto, é quando o profissional deve identificar as informações que serão manipuladas, funções a serem processadas, desempenho desejado, interface, restrições do projeto e critérios de validação, por meio de três etapas especificas: a Análise do Sistema; o Planejamento do Projeto de Software; e a Análise de Requisitos. (MAZZOLA, 199-?).

Já na Fase de Desenvolvimento, será determinado como as funções do software serão realizadas, e define aspectos tais como a arquitetura do software, estruturas dos dados, procedimentos a serem implementados, como transformar o projeto em linguagem de programação, geração de código os procedimentos de teste, sendo organizada por três processos principais: o Projeto de Software; a Codificação; e os Testes de software. (MAZZOLA, 199-?).

E por fim a Fase de Manutenção, iniciada a partir da entrega do software, encarregada de corrigir erros das fases anteriores, inclusão de novas funções, ou adaptação do software a novas configurações de hardware. (MAZZOLA, 199-?).

Assim, tal conceito aplica-se perfeitamente ao presente trabalho, uma vez que será visado um produto final, com a função de resolver um problema do usuário, com eficiência, clareza de funcionalidades e com em menor tempo e custo possíveis, havendo de incluir ao processo pelo menos, as fases descritas acima.

4 TRABALHOS CORRELATOS

Nesta seção serão apresentados alguns trabalhos correlatos que nortearam esta pesquisa.

Araújo, Benevenuto e Golçalves (2013) descrevem uma série de métodos utilizados para análise de sentimentos no Twitter, tendo, entretanto, como foco a língua inglesa. Na investigação é realizada uma análise comparando 8 métodos distintos (emoticons, LIWC, SentiStrength, SenticNet, SASA, Happiness Index e PANAS-t) populares de análise de sentimentos. A análise comparou esses métodos em termos de medida de cobertura e em termos identificação correta de sentimentos. Eles também desenvolveram um novo método que combina abordagens existentes com o objetivo de prover melhores resultados de cobertura de um modo eficiente. O método desenvolvido foi testado por meio de uma ferramenta denominada iFeel, um serviço web que provê uma API aberta para acesso e comparação de resultados por meio de diferentes métodos para um certo texto.

Já Weitzel et al. (2013) também aplicam métodos de análise de sentimentos voltados ao Twitter na língua inglesa. Em seu trabalho, porém, focam na linha do tempo de usuários para aplicarem duas abordagens diferentes, a análise léxica e a análise sintática, retirando os caracteres especiais que sejam dispensáveis com a primeira abordagem, e estruturando o texto para a classificação. Deste modo, tiveram como objetivo comprovar se com a retirada de stop words (palavras sem peso para o significado da frase), haveriam resultados diferentes dos encontrados sem a exclusão das mesmas, e se seria possível relacionar a popularidade dos usuários testados com a polaridade dos tweets. Os autores concluem que não existe diferença significativa na classificação dos sentimentos com a retirada das stop words e que também não há relação entre a popularidade dos usuários com tipo de sentimentos expressos em seus comentários dentro do site.

E por fim, Lopes (2009), com a ferramenta “opsys” que tem como principal finalidade a classificação de opiniões encontradas na web com foco nas empresas. A empresa deve se cadastrar para poder fazer uma avaliação de sua marca e a ferramenta busca, de forma automática, opiniões sobre ela em intervalos regulares, e ao longo do tempo fazendo um histórico com as avaliações dos produtos e da marca, permitindo ao empresário ter uma noção mais clara sobre a evolução da

empresa. Baseando em técnicas do Processamento de Linguagem Natural, e com estrutura similar a do presente trabalho que poderá ser visto a partir da metodologia descrita no próximo tópico.

5 METODOLOGIA

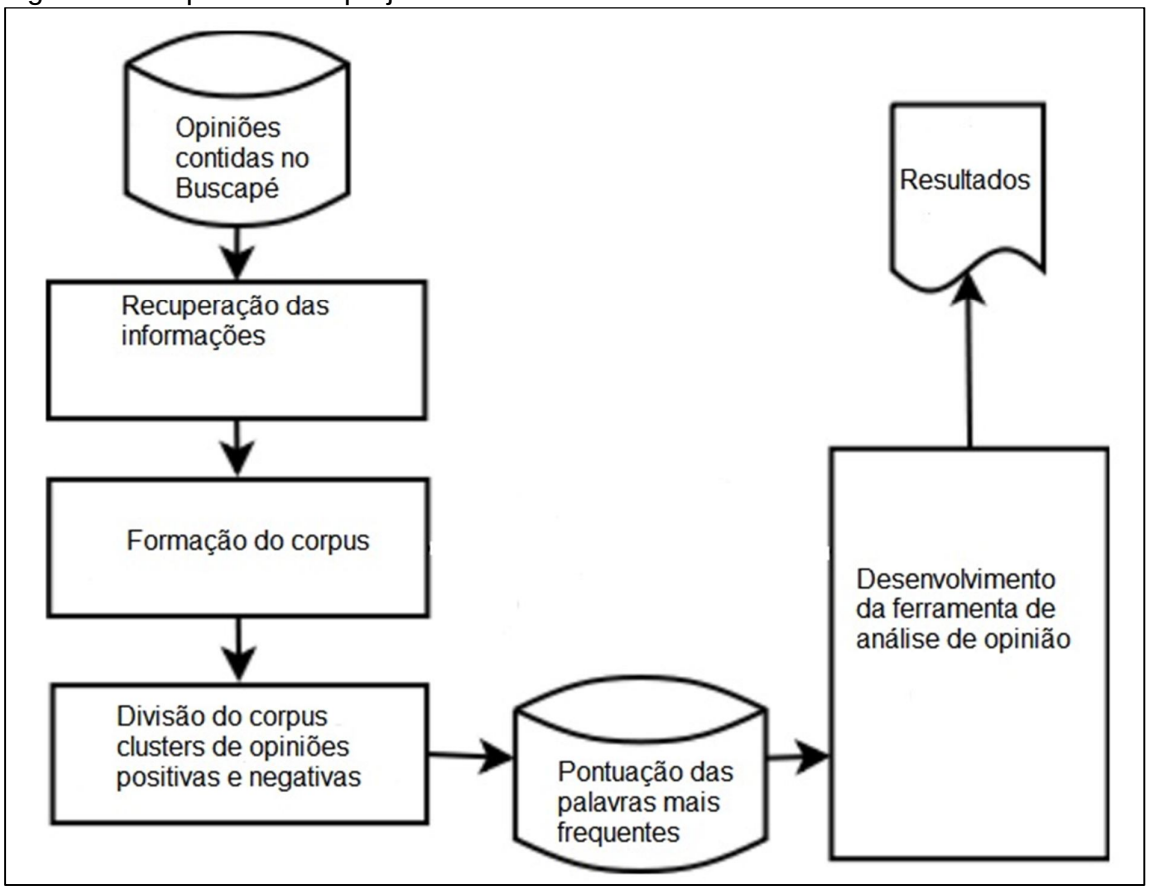
Segundo Gil (2010), as pesquisas exploratórias têm como propósito proporcionar maior familiaridade com o problema, com vistas a torna-lo mais explícito ou a construir hipóteses. Seu planejamento tende a ser bastante flexível, pois interessa considerar os mais variados aspectos relativos ao fato ou fenômeno estudado. O autor também afirma que a maioria das pesquisas realizadas com propósitos acadêmicos, pelo menos num primeiro momento, assumem o caráter de pesquisa exploratória, pois neste momento é pouco provável que o pesquisador tenha uma definição clara do que irá investigar.

Portanto, este projeto é inicialmente uma pesquisa exploratória, pois estudou e implementou uma abordagem que possibilitou o desenvolvimento de um protótipo de sistema de análise de sentimentos para a língua portuguesa. Basicamente, para cumprimento de tal objetivo foram necessários dois passos: a definição da arquitetura e funcionamento do sistema e o processo de formação de um corpus linguístico. Tais passos são descritos nas seções seguintes.

5.1 ARQUITETURA E FUNCIONAMENTO DO SISTEMA PROPOSTO

Tendo em vista a recuperação e manipulação das opiniões citadas anteriormente, o presente trabalho foi dividido basicamente em cinco processos conforme a Figura 3 ilustra.

Figura 8 – Arquitetura do projeto.



Fonte: Elaborada pelo autor.

Assim como ilustrado na Figura 3, para a elaboração do protótipo do sistema de análise de sentimento, em primeiro lugar houve a recuperação das opiniões contidas dentro do site do “Buscapé”, tal processo ocorreu de maneira automatizada, o que possibilitou a formação do corpus linguístico, que foi o resultado da união de todas as opiniões obtidas a partir da recuperação. Esse corpus foi dividido em dois clusters com a abordagem de Bag-Of-Words, um contendo as opiniões positivas e outro contendo as negativas. Posteriormente, tais clusters foram submetidos à ferramenta de análise de corpus, o “Corpógrafo”, que indicou a frequência de unigramas[4], bigramas e trigramas mais significativos, que puderam, assim, ser pontuados (recebendo um peso), visando criar um ranking de palavras indicativas de opinião que foram utilizadas no processo de análise de sentimentos. Para esta investigação foi considerado que a palavra/expressão que possuísse a maior

[4] Os unigramas são palavras únicas, já os bigramas e trigramas são, respectivamente, sequências de duas ou três palavras.

frequência dentro de um cluster positivo ou negativo, receberia o peso máximo (1 para positivo e -1 para negativo); e as demais receberam pesos proporcionais à sua frequência no corpus. A utilização de pesos entre -1 e 1 foi escolhida por ser algo comum em sistemas de processamento de língua natural, os quais utilizam frequências relativas no processo de ponderação de textos. As expressões repetidas entre ambos tiveram seus pesos somados. Tal pontuação de expressões serviu como base para o analisador de sentimentos, que comparou as palavras e expressões da opinião recebida com as de sua base, atribuindo o peso pré-estabelecido no processo anterior, tornando, assim, possível a soma de todos esses pesos, possibilitando através do resultado final, determinar se tal opinião é positiva ou negativa. Ao fim deste processo, opiniões cujas pontuações (soma dos pesos recebidos) estejam acima de zero, provavelmente indicam um caráter positivo, já aquelas abaixo de zero tem um caráter negativo.

Para exemplificar o método descrito acima, suponha uma opinião sobre um smartphone como "*É um ótimo smartphone, não trava, tem uma câmera boa, apesar do preço*", obtida de um site da web. Essa opinião foi submetida à aplicação, que tem em sua base de conhecimento as expressões (unigramas, bigramas e trigramas) já com seus devidos pesos pré-estabelecidos, assim, cada expressão contida na opinião foi comparada com as da base. Para o exemplo dado, as expressões "smartphone" e "boa" foram encontradas dentro da base de conhecimento. Deste modo, estas expressões foram pontuadas conforme os valores agregados juntamente com as expressões, obtendo-se então como peso "smartphone" com peso positivo de 0.31, e negativo de -0,06, e a palavra "boa" obteve peso 0.17. A partir daí as expressões foram trabalhadas apenas por seus pesos para determinar a polaridade da opinião, onde opiniões com peso final menor que zero são consideradas negativas, e acima de zero são consideradas positivas. Para o exemplo a soma da opinião foi: 0.31 + (-0.06) + 0.17= 0.42. Como resultado final, a opinião obteve peso 0.42 e, assim, é classificada com polaridade positiva.

Após a análise, os resultados indicados pelo analisador foram avaliados para viabilizar a conclusão ao final do presente projeto. Para isso foi considerada a medida de precisão. A precisão indica qual o percentual de acerto do sistema, considerando o número de opiniões classificadas corretamente em positivas ou negativas.

Nos próximos tópicos as etapas referentes à construção do sistema serão detalhadas.

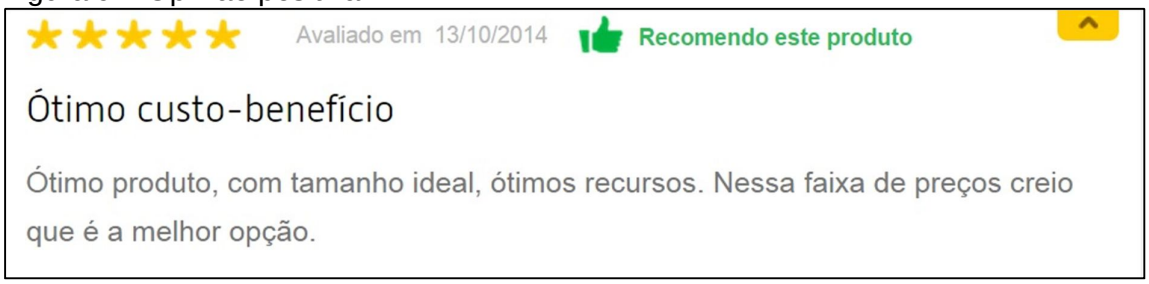
5.2 CRIAÇÃO DE UM CORPUS DE OPINIÕES

Como abordado na seção 2, o presente trabalho teve como principal objetivo o desenvolvimento de um protótipo de sistema de análise de sentimentos para a definição automática de opiniões encontradas na web.

Para tanto foi utilizado como repositório de opiniões o site do “Buscapé” onde se encontram listas com opiniões positivas e negativas acerca de produtos específicos pesquisados.

As Figuras 4 e 5 mostram exemplos de comentários, tanto positivo quanto negativo, que são encontrados no site e que foram utilizados no processo de mineração de opinião. Tais comentários serviram como base para a montagem de um corpus linguístico utilizado no treinamento de um sistema de análise de sentimentos.

Figura 9 – Opinião positiva.

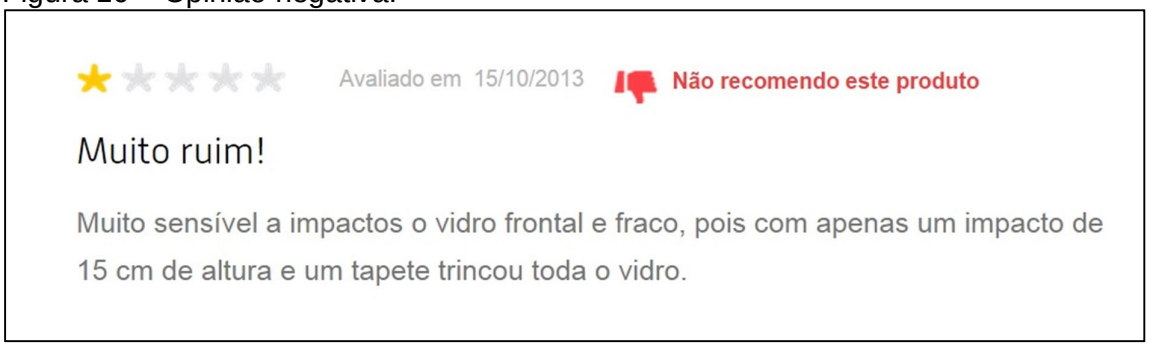


Fonte: Smartphone... (2014).
Nota: Adaptada pelo autor.

A Figura 4 como se pode perceber, mostra uma opinião positiva com relação a um produto específico, neste caso um smartphone. Observa-se que o usuário, além de escrever os comentários que julga pertinente, também faz uma avaliação por meio de estrelas. Em tal avaliação quanto maior o número de estrelas, melhor julgado o produto e quanto menor, pior avaliado. Há também um item de recomendação, sendo ela positiva ou negativa, juntamente com uma listagem de prós e contras pré-definidos no site.

Com uma leitura rápida já é possível encontrar elementos em comum dentro do texto escrito como, por exemplo, a palavra “ótimo”, encontrada em duas vezes na opinião explicitada na Figura 4.

Figura 10 – Opinião negativa.



Fonte: Smartphone... (2014).
Nota: Adaptado pelo autor.

Já na Figura 5, em contraponto à Figura 4, é explicitada uma opinião negativa em relação a outro smartphone, sendo possível notar o menor número de estrelas na avaliação, o que indica uma recomendação negativa do produto.

Deste modo, para haver a criação do presente corpus, foram utilizadas cerca de cinco mil opiniões semelhantes aos exemplos destacados anteriormente. Entretanto, não foi adotado nenhum critério de balanceamento no número de opiniões positivas e negativas. Captadas através de uma extensão do navegador “google chrome” chamada “Web Scraper”, que tem o papel de capturar de forma automática o conteúdo de determinadas divisões dentro de uma página indicada, retornando um arquivo tipo CSV para cada produto em específico. Tais arquivos foram posteriormente reunidos em um único arquivo TXT para opiniões positivas e para as opiniões negativas.

Tal processo foi aplicado a opiniões sobre nove smartphones diferentes, sendo eles o Samsung Galaxy Notes 3, Samsung Galaxy S 5, Samsung Galaxy S 4, Samsung Galaxy S 3, Iphone 5s, Lg G3, Lg Nexus 5, Motorola Moto X e Sony Xperia Z 2. Tais modelos foram escolhidos por conta da sua popularidade no mercado.

Após a união de todas as opiniões em um único arquivo de texto para cada polaridade, tais arquivos foram submetidos a um processo de pré-processamento para diminuir a probabilidade de conflitos referentes à acentuação, caracteres especiais e entre letras maiúsculas e minúsculas. Além disso, também foram retiradas as stopwords cujo a lista pode ser vista através do APÊNDICE A deste

trabalho, que são palavras que não adicionam sentido nenhum dentro qualquer possível contexto. Tal processo se deu de forma automatizada por meio de uma aplicação criada especificamente para este fim, utilizando-se a linguagem de programação PHP.

Este processamento tornou possível a submissão de tais textos à ferramenta do corpógrafo, que através da contagem de frequência, identificou as palavras e expressões mais relevantes dentro de cada polaridade do corpus formado. Assim, foi possível identificar os unigramas, bigramas e trigramas que mais foram frequentes dentre todas as opiniões recuperadas nos processos anteriores.

Destaca-se que optou-se por uma filtragem de termos realizada de modo manual, para que houvesse maior controle dos dados que seriam utilizados posteriormente, priorizando a qualidade dos dados. Para a lista de unigramas, por se tratar de palavras únicas, a ferramenta acabou por listar todas as palavras contidas no corpus, deste modo definiu-se que para o presente projeto seriam utilizadas em torno das cinquenta mais frequentes, com exceção das palavras que denotam marcas como Samsung, Apple ou Sony, e também palavras relacionadas a nomes de modelos de smartphones como Iphone, Moto X ou Nexus.

Com relação aos bigramas e, principalmente aos trigramas, este processo se deu de maneira a serem escolhidas as expressões que fariam sentido e que representassem a real ideia de seu corpus como, por exemplo, ao se analisar os bigramas positivos foram desconsideradas expressões como “não gostei” ou “deixa desejar”. Durante a análise dos termos negativos, expressões como “muito bom” e “recomendo produto” foram descartadas, além de haver o mesmo critério de descartar marcas e nomes de modelos utilizado nos unigramas, sendo aproveitadas todas as expressões consideradas relevantes.

Assim, ao final de todo este processo, foram consideradas para a aplicação um total de 98 unigramas, sendo 50 positivos e 48 negativos, 70 bigramas, 48 positivos e 22 negativos e 32 trigramas, sendo 29 positivos e 3 negativos.

Finalizada a etapa de análise das expressões que seriam utilizadas no processo de pontuação pela aplicação, foram estabelecidos os pesos que seriam atribuídos a cada expressão. Primeiramente, como citado anteriormente, definiu-se que as expressões teriam peso entre 1 e -1. Para isso, todas as expressões mais frequentes tiveram atribuídos a elas peso 1 positivo, se fosse um grupo de expressões positivas(unigramas positivos, bigramas positivos e trigramas positivos)

e peso -1 se fossem dos grupos negativos. Deste modo, para as expressões subsequentes, seu peso foi definido pela divisão de sua frequência, pela frequência da expressão com peso 1, por exemplo, para os bigramas positivos, a expressão mais frequente foi “muito bom”, ocorrendo 697 vezes dentro do corpus, e a expressão “ótimo produto” ficou em quinto lugar com 174 ocorrências. Desta maneira, para se definir o peso da expressão “ótimo produto” dividiu-se 174 por 697, o que lhe gerou de peso final 0,24. Este procedimento foi aplicado para todas as outras expressões obtidas nas etapas anteriores, gerando o peso de cada uma delas.

Na seção seguinte, será abordado o processo de desenvolvimento do software para a análise de sentimentos, que faz uso das expressões e pesos definidos nesta fase.

5.3 DESENVOLVIMENTO DO SOFTWARE

Como abordado no início da seção anterior, o presente trabalho teve como principal objetivo o desenvolvimento de um protótipo de sistema de análise de sentimentos para a definição automática de opiniões encontradas na web. Deste modo, a presente seção se desenvolve com o objetivo de explicar o funcionamento do software desenvolvido.

Primeiramente, para possibilitar o desenvolvimento do software, todas as expressões selecionadas e pontuadas, conforme descritas na seção 5.2, foram inseridas dentro de um banco de dados MySQL, tal escolha para o sistema de gerenciamento de banco de dados se deve ao fato de ser um sistema gratuito, além de ter grande compatibilidade com a linguagem de programação adotada para o software, o PHP.

As opiniões foram dispostas em seis tabelas, com o intuito de facilitar a manipulação futura das expressões inseridas no banco, sendo elas as tabelas de unigramas, bigramas e trigramas divididos em positivos e negativos.

Para o software a linguagem de programação escolhida para o desenvolvimento foi o PHP, devido às próprias características da linguagem, tais como: suporte à aplicações web, sistema multiplataforma, suporte a um grande número de banco de dados além de possuir seu código fonte aberto.

Assim, o programa foi desenvolvido com o objetivo de analisar o sentimento expresso dentro de uma opinião qualquer, neste caso sobre smartphones (uma vez que o corpus no qual ele se baseia está relacionado a este tipo de produto).

O software é constituído de quatro fases simples, a entrada de dados via interface, como ilustrado na Figura 6, o pré-processamento do texto de entrada, a análise e o retorno, mostrados na Figura 7.

Figura 6 – Interface do Software.

Fonte: Elaborada pelo autor.

Como pode ser observado a partir da Figura 6, a interface do programa é muito simples, constituída de uma área para a inserção do texto e dois botões, um visando a classificação, e outro para limpar a área de texto, além do conteúdo textual indicando a função do programa. Nesta área o usuário deverá inserir sua opinião a fim de classificá-la.

Após a ativação do botão “classificar”, o programa entra em um estado de processamento que não está visível para o usuário.

Nesta fase, primeiramente o texto inserido passará por um pré-processamento, onde a opinião inserida terá um tratamento idêntico ao dado para o corpus anteriormente, exatamente com o propósito de deixar os textos que serão comparados em formatos semelhantes, sem acentos, caracteres especiais e sem as stopwords, minimizando, assim, as possibilidades de conflito.

Já com o texto pré-processado, esta opinião será disposta em um vetor, como uma Bag-of-words, que será comparado aos vetores recuperados do banco de

dados com as expressões do corpus na fase de análise. Cabe destacar que optou-se inicialmente por utilizar de forma concomitante, tanto os unigramas quanto bigramas e trigramas para a classificação da opinião. A hipótese era de que os três de forma conjunta seriam capazes de classificar satisfatoriamente, uma opinião.

Para a pontuação da opinião, o vetor formado será percorrido, e comparado aos vetores recuperados das tabelas inseridas no banco de dados, primeiramente aos trigramas, bigramas e unigramas (nesta ordem) positivos, e depois pelo mesmo processo para os negativos. A comparação consiste em confrontar as expressões, sendo que se ela estiver contida no vetor recuperado do banco, a opinião receberá a pontuação dada à expressão anteriormente, somando-se assim à que já havia sido atribuída.

Ao final, somam-se os resultados obtidos a partir das expressões positivas e negativas, e a que obtiver uma pontuação maior, definirá a polaridade da opinião, ou seja, se ela é positiva ou negativa.

Após o fim da fase de análise e com a opinião classificada, o programa mostrará o resultado obtido ao usuário por meio de um alerta, como na Figura 7.

Figura 7 – Retorno do Software.

Fonte: Elaborada pelo autor.

Para a avaliação do sistema, as etapas de interface e retorno foram retiradas do processamento, para possibilitar que o processo fosse feito de forma automatizada. Tal avaliação será descrita de maneira mais aprofundada na próxima seção.

6 RESULTADOS

Ao fim do desenvolvimento, a aplicação foi submetida a testes de precisão que visaram verificar a sua eficiência na tarefa de classificação. Os testes também foram definidos de forma a verificar se a hipótese inicial de utilizar de forma conjunta, tanto os unigramas quanto bigramas e trigramas para a classificação da opinião traria, de fato, os melhores resultados, ou se haveria alguma outra combinação que pudesse melhorar a precisão do sistema.

Para esta verificação, um corpus de teste foi formado com opiniões disponíveis em vários sites e-commerce na internet, tais como “americanas.com.br”, “shoptime.com.br”, dentre outros. Visando uma maior confiabilidade, não foram utilizadas opiniões contidas no site do “Buscapé”, uma vez que o mesmo serviu como repositório do corpus de treinamento da aplicação.

No total o corpus de teste foi formado por 100 opiniões, sendo 50 positivas e 50 negativas. O sistema foi testado em diferentes situações, avaliando-se sempre a porcentagem de acertos (precisão) em cada uma das situações propostas.

No total a aplicação foi testada de oito maneiras diferentes, sendo elas:

- Teste 1:utilizando todo o repositório formado ao logo do processo, portanto, unigramas, bigramas e trigramas positivos e negativos;
- Teste 2: utilizando a tabela de unigramas negativos modificada e sem a palavra “não” inclusa, uma vez que observou-se que houve grande diferença de frequência entre a palavra “não” e as demais, sendo importante avaliar a sua influência no resultado;
- Teste 3: não foram retiradas as “stopwords” na etapa de pré-processamento da opinião;
- Teste 4: a opinião foi confrontada apenas com as tabelas de unigramas;
- Teste 5:apenas os bigramas foram comparados à opinião;
- Teste 6: foram unidos os unigramas aos bigramas;
- Teste 7: uniu os unigramas e os trigramas
- Teste 8 que testou a união dos bigramas com os trigramas,

O quadro apresentado na Figura 8 traz um resumo das configurações testadas.

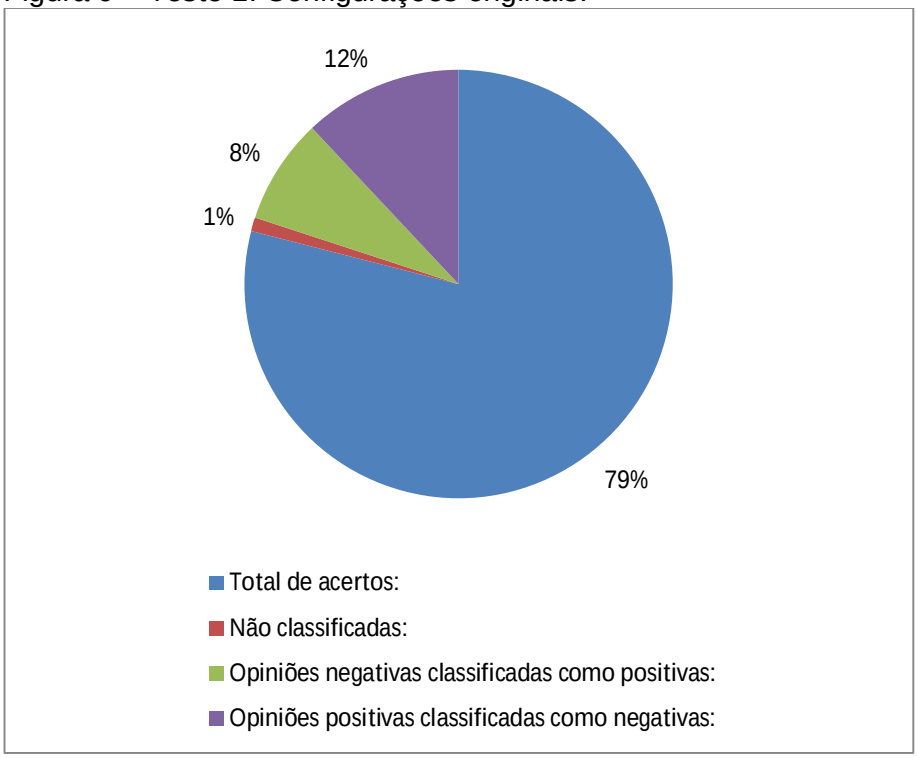
Figura 8 – Resumo das configurações dos testes.

	Palavra “não”	Retirando Stopwords	Unigramas	Bigramas	Trigramas
Teste 1	X	X	X	X	X
Teste 2		X	X	X	X
Teste 3	X		X	X	X
Teste 4	X	X	X		
Teste 5	X	X		X	
Teste 6	X	X	X	X	
Teste 7	X	X	X		X
Teste 8	X	X		X	X

Fonte: Elaborada pelo autor.

Todos os testes supracitados foram comparados às mesmas opiniões, de forma automatizada, uma vez que todas as opiniões colhidas para testes foram inseridas em uma nova tabela no banco de dados, juntamente com a sua polaridade correta. Deste modo um contador era incrementado toda vez em que ocorriam avaliações corretas, e outro quando erros foram cometidos, além de também identificar o número de opiniões positivas e negativas identificadas corretamente, possibilitando uma avaliação quantitativa do sistema.

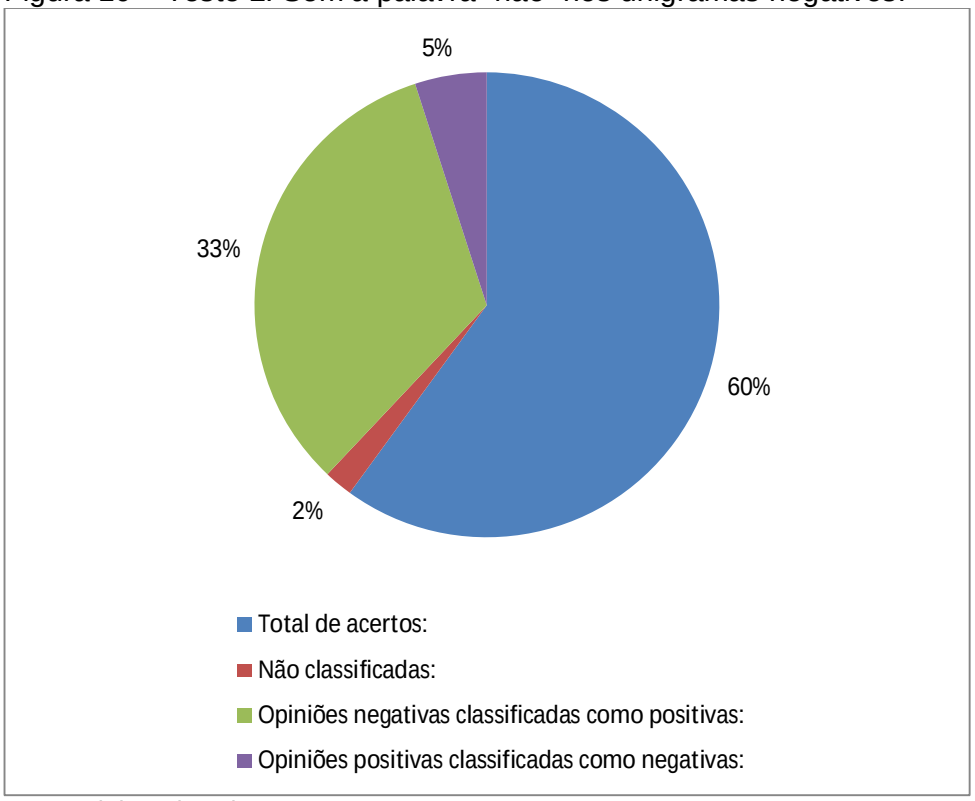
Figura 9 – Teste 1: Configurações originais.



Fonte: Elaborada pelo autor.

Ao final do Teste 1 (utilizando todo o repositório), foi computado um total de 79% de acerto do sistema em relação à polaridade da opinião, sendo que 8 opiniões negativas foram classificadas erroneamente como positivas, 12 opiniões positivas foram classificadas como negativas e uma opinião não foi classificada, assim como apresentado no gráfico da Figura 9. Tais resultados mostram que o sistema obteve boa taxa de precisão, uma vez que obteve quase 80% de opiniões corretamente classificadas. Dentre seus erros, observa-se uma pequena taxa de incapacidade de classificação (apenas 1%), e um número maior de opiniões positivas classificadas como negativas do que o contrário, o que pode ser um indício de um possível erro quanto aos pesos atribuídos às expressões negativas, uma vez que seu repositório de opiniões foi significativamente menor do que o de opiniões positivas.

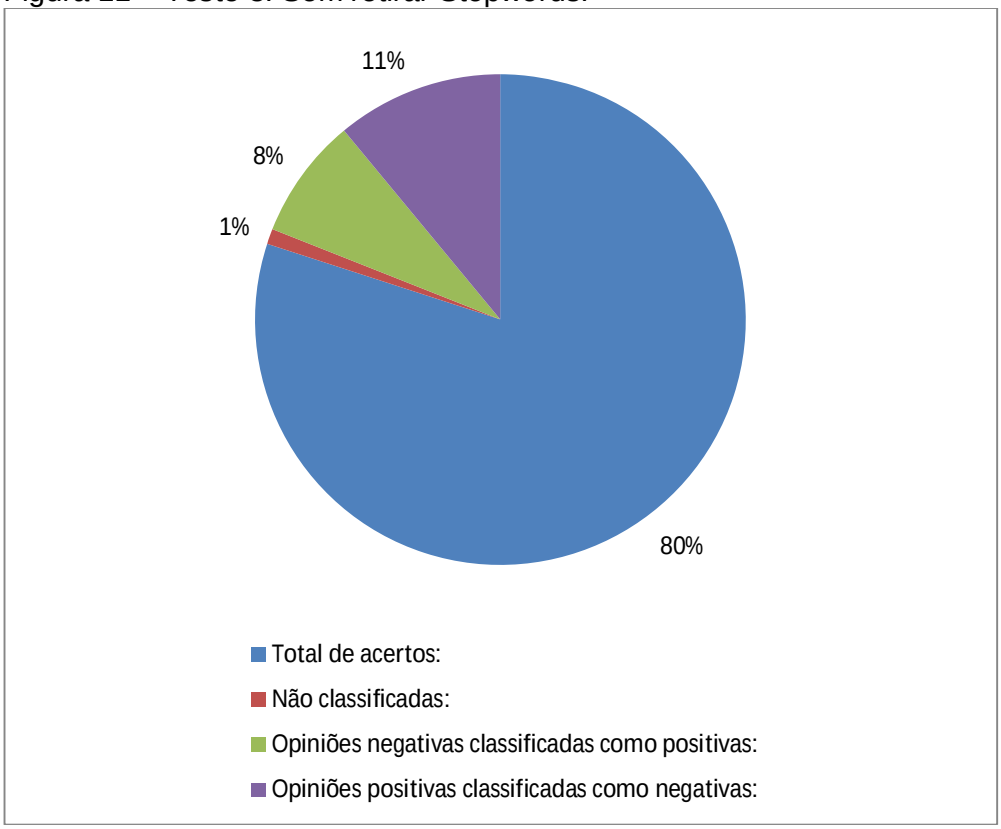
Figura 10 – Teste 2: Sem a palavra “não” nos unigramas negativos.



Fonte: Elaborada pelo autor.

O Teste 2 que utilizou-se de uma tabela de pesos sem a inclusão da palavra “não” dentre os unigramas negativos, obteve uma taxa de precisão de 60%, sendo que 33 opiniões negativas foram consideradas positivas e o contrário aconteceu 5 vezes, além de duas opiniões não classificadas, conforme a Figura 10. Deste modo, houve um decréscimo da precisão em relação aos testes que utilizaram unigramas acrescidos da palavra “não”. Sua grande taxa de opiniões negativas classificadas como positivas, sugere um grande desequilíbrio no número de expressões reconhecidas como positivas em relação às negativas, uma vez que a palavra mais frequente dentre as opiniões negativas foi descartada.

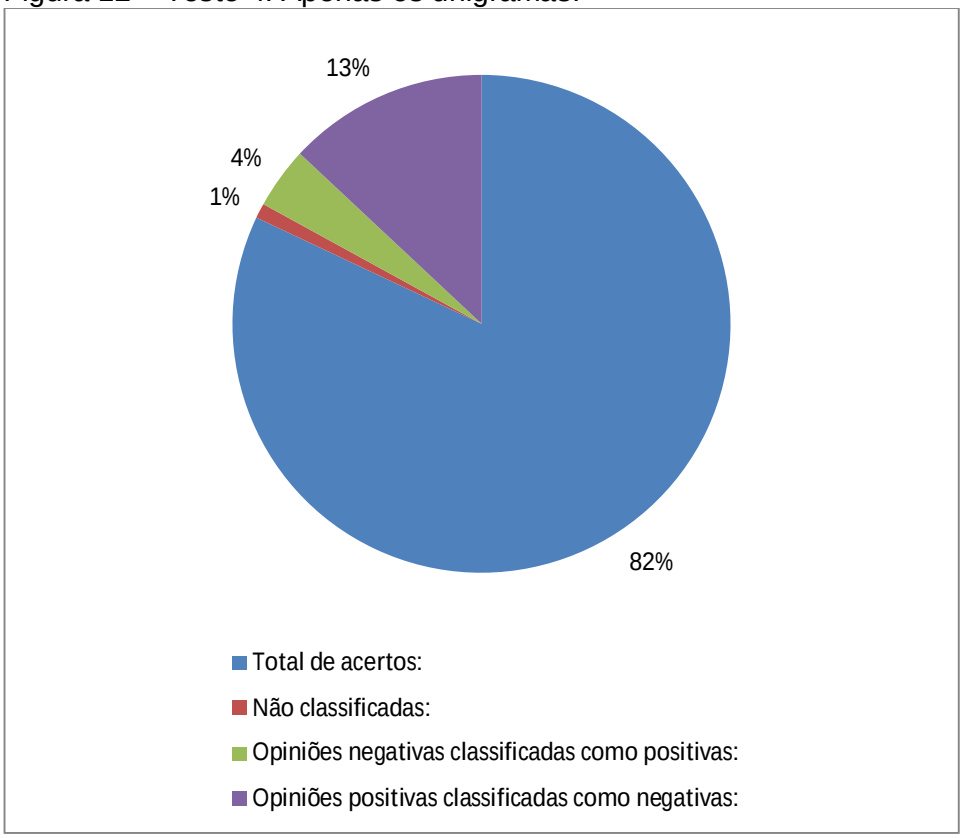
Figura 11 – Teste 3: Sem retirar Stopwords.



Fonte: Elaborada pelo autor.

No Teste 3 em que as “stopwords” não foram retiradas da opinião submetida ao sistema, a taxa de precisão obteve um total de 80% de acertos, sendo que dentre as classificadas erroneamente, ocorreram 8 erros em relação as opiniões negativas (classificadas como positivas), e 11 em relação as opiniões positivas, além de uma delas não ter sido classificada, de acordo com a Figura 11. Esta configuração teve um aumento ínfimo de precisão em relação à primeira configuração (teste 1), não sendo possível indicar com certeza que a permanência das stopwords dentro da opinião submetida ao teste torne o sistema mais preciso.

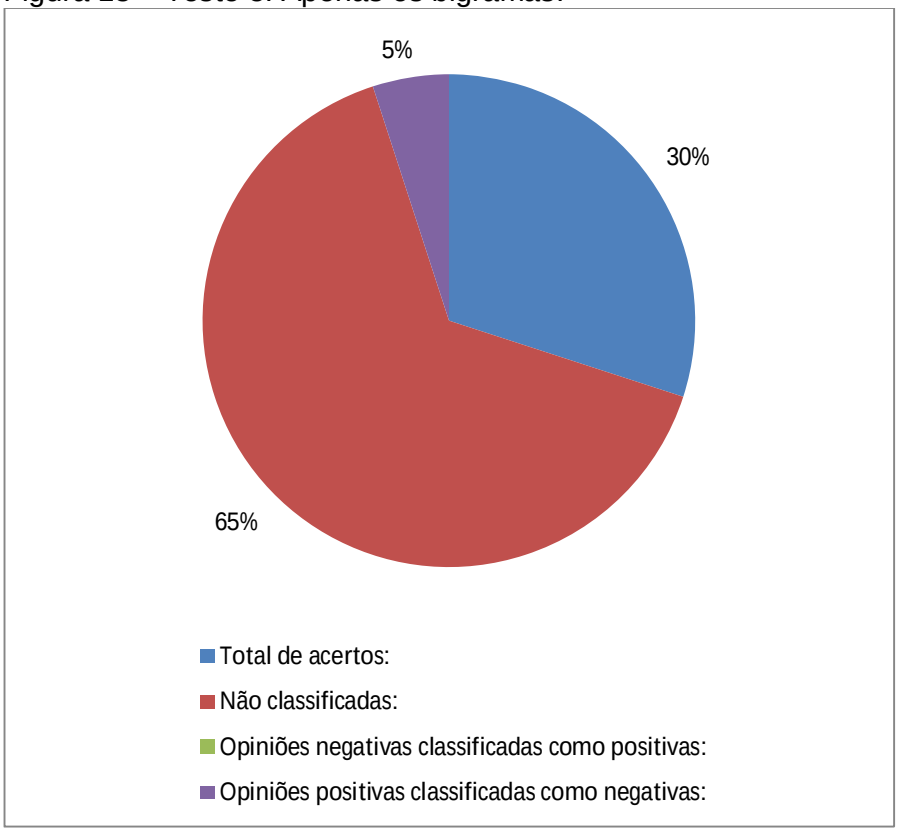
Figura 12 – Teste 4: Apenas os unigramas.



Fonte: Elaborada pelo autor.

Após o Teste 4, no qual foram considerados apenas os unigramas para a comparação com a opinião submetida, o sistema mostrou-se ligeiramente mais preciso, com uma taxa de acerto de 82%, com uma opinião não classificada, além de 4 opiniões negativas classificadas como positivas, e 13 positivas classificadas como negativas, conforme indicado na Figura 12. Apesar de um ligeiro aumento na taxa de opiniões negativas classificadas como positivas, a taxa de opiniões negativas classificadas como positivas caiu pela metade com esta configuração, o que pode ser devido um maior equilíbrio entre o número de expressões negativas e positivas reconhecidas no processo. Isso se deve ao fato haver um número parecido de unigramas positivos e negativos o que, por sua vez, diminuiu a influência de um número de opiniões negativas coletadas para treinamento ser bem menor do que o número de opiniões positivas.

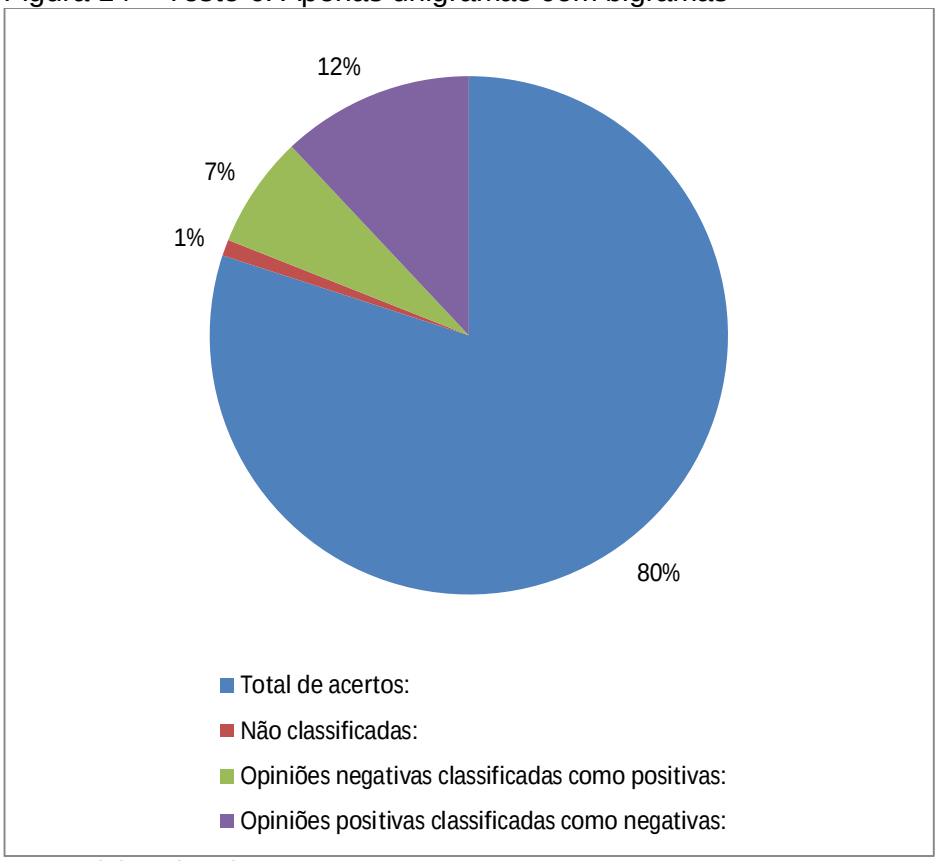
Figura 13 – Teste 5: Apenas os bigramas.



Fonte: Elaborada pelo autor.

A Figura 13 ilustra os resultados obtidos a partir do Teste 5, no qual apenas as expressões dentro dos bigramas foram comparadas às opiniões do corpus de teste. Observa-se uma queda no número de opiniões classificadas corretamente, com uma taxa de precisão de 30%, e com 65% das opiniões não sendo classificadas, além de 5 opiniões positivas serem classificadas como negativas. Tal queda de precisão na classificação das opiniões pode ser decorrente da diminuição do número de expressões deste tipo em relação aos unigramas, e da maior dificuldade em se ocorrer este tipo de expressão, fazendo com que a maior parte das opiniões não fosse sequer classificada com alguma polaridade.

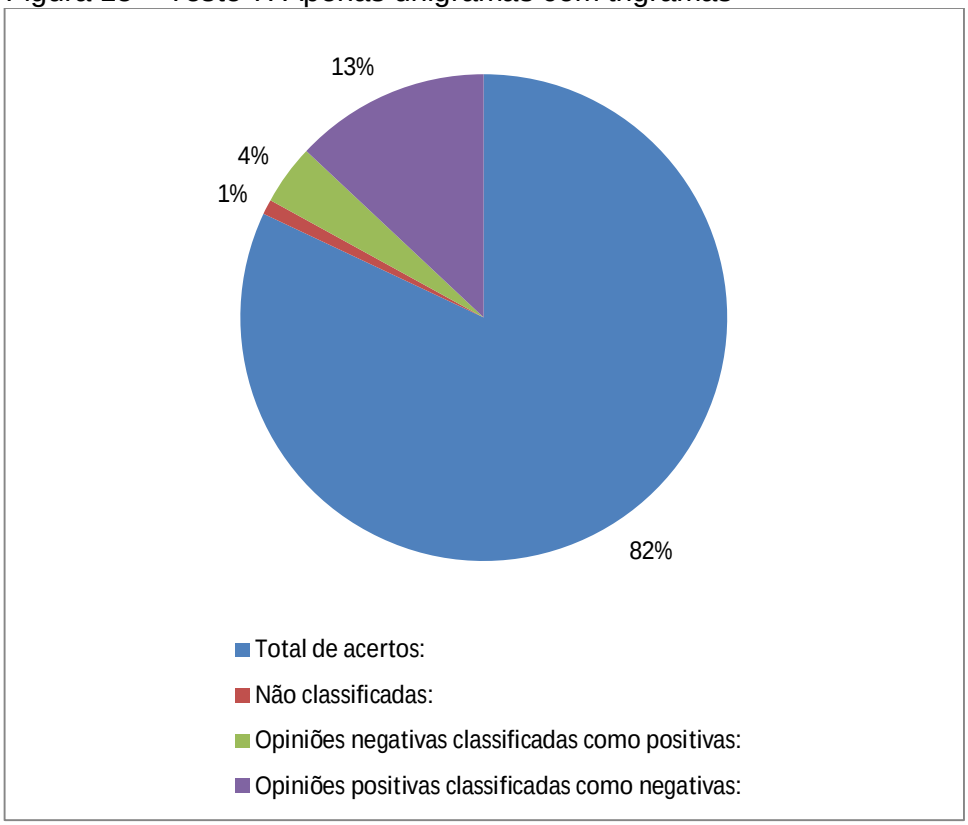
Figura 14 – Teste 6: Apenas unigramas com bigramas



Fonte: Elaborada pelo autor.

O sexto teste, como apresentado no gráfico da Figura 14, uniu os unigramas aos bigramas para a classificação da opinião e obteve uma alta taxa de precisão, com 80% de acerto e sete opiniões negativas classificadas como positivas, 12 positivas classificadas como negativas e uma opinião não classificada. Porém se comparados esses resultados com o do Teste 4 no qual apenas os unigramas foram considerados, este teste teve uma queda de precisão, o que sugere que a adição dos bigramas foi prejudicial à taxa de precisão do software, isso pode ser decorrência do desequilíbrio encontrado no número de expressões dos bigramas positivos e negativos.

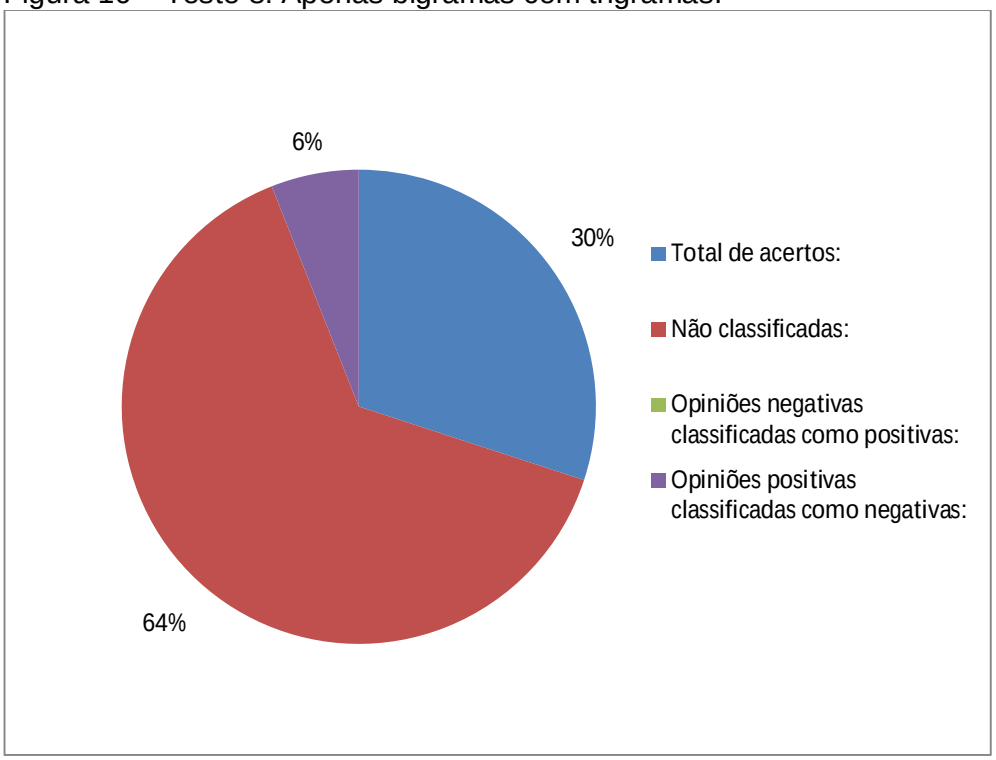
Figura 15 – Teste 7: Apenas unigramas com trigramas



Fonte: Elaborada pelo autor.

O gráfico de resultados apresentado pelo Teste 7 (Figura 15), que incluiu na classificação das opiniões os unigramas juntamente aos trigramas, é idêntico ao gráfico apresentado ao final do Teste 4, que utilizou-se apenas dos unigramas para a classificação. Tal resultado indica que a inclusão dos trigramas juntamente com os unigramas não influenciou os resultados obtidos, em relação aos testes unicamente com os unigramas, sugerindo que os trigramas acabaram por ser pouco relevantes para as classificações. Isso pode se dar ao fato de a possibilidade de união de 3 palavras em determinada ordem ocorrer com uma probabilidade pequena, o que acarretou em um número pequeno de trigramas significativos serem encontrados para o corpus da aplicação e, consequentemente, diminuindo a relevância dos trigramas no sistema.

Figura 16 – Teste 8: Apenas bigramas com trigramas.



Fonte: Elaborada pelo autor.

Por fim, o último teste realizado que uniu os bigramas aos trigramas, assim como indicado pelo gráfico da Figura 16, indica resultados muito semelhantes aos do Teste 5 que utilizou apenas os bigramas, com a única diferença, que uma opinião que não foi classificada pelo Teste 5 passou a ser classificada erroneamente como negativa. Apesar de ser uma diferença pequena, este teste mostra que os trigramas influenciam a classificação feita pelo sistema, porém ainda é necessário uma melhora em seu conjunto de expressões relevantes.

A tabela ilustrada na Figura 17 mostra resumidamente todos os testes realizados para o sistema, dando uma visão geral dos resultados obtidos.

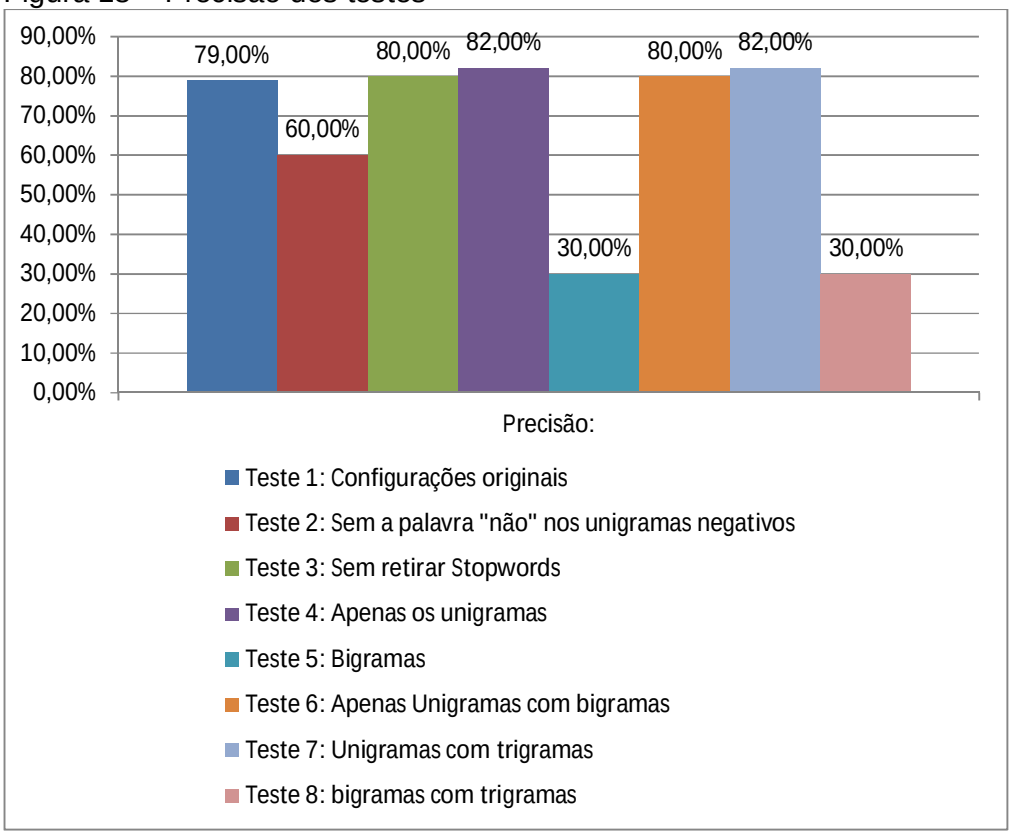
Figura 17 – Resumo dos resultados.

	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5	Teste 6	Teste 7	Teste 8
Total de acertos:	79	60	80	82	30	80	82	30
Total de erros:	20	38	19	17	5	19	17	6
Não classificadas:	1	2	1	1	65	1	1	64
Opiniões negativas classificadas como positivas:	8	33	8	4	0	7	4	0
Opiniões positivas classificadas como negativas:	12	5	11	13	5	12	13	6
Precisão:	79,00%	60,00%	80,00%	82,00%	30,00%	80,00%	82,00%	30,00%

Fonte: Elaborada pelo autor.

Ao final de todos os testes, foi possível notar a importância dos unigramas dentro do sistema que utiliza o método de classificação proposto nesta investigação, uma vez que eles potencializam a taxa de precisão quando utilizados sozinhos e fazendo-a diminuir quando não utilizados, como pode ser visto a partir da Figura 18. Tal ocorrência se deve ao fato unigramas serem mais facilmente reconhecidos dentro de um texto, uma vez que são palavras únicas e no presente projeto estarem em maior número do que os outros tipos de expressões.

Figura 18 – Precisão dos testes



Fonte: Elaborada pelo autor.

Também pode ser observado na Figura 18 que a frequência da palavra “não” dentro dos unigramas negativos é de grande importância para uma classificação mais precisa das opiniões, além de uma baixa representatividade dos trigramas para o sistema.

7 CONSIDERAÇÕES FINAIS

Os resultados obtidos a partir do presente trabalho contribuem para pesquisas no campo do processamento de linguagem natural em língua portuguesa, uma vez que o método proposto obteve uma taxa de precisão de 82% na classificação das opiniões submetidas ao sistema, o que pode ser considerado como uma abordagem com um bom potencial.

Porém o presente método também apresentou limitações, como as indicadas durante a discussão dos resultados, como a deficiência do sistema em classificar as opiniões corretamente sem a inclusão dos unigramas. Isso decorre, provavelmente, do fato de haverem poucas opções de expressões dentre bigramas e trigramas. Como possibilidade de extensão desta pesquisa, sugere-se um foco maior na coleta de mais expressões deste tipo, o que pode gerar melhores resultados.

Outra limitação encontrada refere-se ao processo de coleta do repositório de opiniões, em que o número de opiniões positivas encontradas foi muito maior do que o número de negativas, o que afetou a diversidade e número de expressões negativas utilizadas no sistema. Para trabalhos futuros com a mesma temática sugere-se a utilização de outro repositório de opiniões, mais consistente, balanceado e com número maior de opiniões negativas em seu escopo. Também poderá ser realizada por especialistas uma coleta mais seletiva das opiniões utilizadas para o corpus, que pode melhorar a qualidade do repositório.

Deste modo, conclui-se que o método proposto e implementado por meio de um software, apesar de simples, tem um grande potencial, já que se alcançaram bons índices de precisão ao classificar opiniões, sendo pouco custoso já que trabalha apenas no nível superficial do texto. Deste modo, o presente trabalho contribui para pesquisas ligadas ao processamento de linguagem natural, principalmente em língua portuguesa, uma vez que há poucos trabalhos específicos nesta língua. Porém ainda há muito a ser melhorado e novas propostas podem contribuir ainda mais para o crescimento de pesquisas deste tipo.

REFERÊNCIAS

AFONSO, D.; GUEDES, R.; MAGALHÃES, L. H. de. Mineração de Opiniões de Usuários na Busca de Conhecimento, **Revista das Faculdades Integradas Vianna Júnior,** Juiz de Fora, out. 2010. Disponível em: <http://www.viannajr.edu.br/files/uploads/20131001_141137.pdf> Acesso em: 08 maio 2014.

ARAÚJO, M.; BENEVENUTO, F.; GONÇALVES, P. **Métodos para Análise de Sentimentos no Twitter**. In: Simpósio Brasileiro de Sistemas Multimídia e *Web (WEBMEDIA).* Salvador, nov, 2013. Disponível em: <http://homepages.dcc.ufmg.br/~fabricio/download/webmedia13.pdf> Acesso em: 18 maio 2014.

ARRUDA, C; PENIDO, E. ROSSI, A. **BuscaPé: do empreendedorismo à inovação aberta**. Casos FDC, 2011. Disponível em: <http://acervo.ci.fdc.org.br/AcervoDigital/Casos/Casos%202010/CF1005.pdf > Acesso em 13 abr. 2014.

BAISH, L. V. et al**. A análise do perfil do cliente como estratégia competitiva em uma escola de idiomas de Santa Maria - RS**. In: SLADE BRASIL/2006 ENCONTRO LUSO-BRASILEIRO DE ESTRATÉGIA, Balneário Camboriú, 2006. Disponível em: <http://www.ead.fea.usp.br/eadonline/grupodepesquisa/publica%C3%A7%C3%B5es/rolando/47.htm >. Acesso em: 25 mar. 2014.

BARROS, F.; LIMA, D.; SILVA, N. R. SAPair: Um Processo de Análise de Sentimento no Nível de Característica, **labic.icmc.usp.br**, 2012. Disponível em: <http://www.labic.icmc.usp.br/wti2012/artigos/105283.pdf>Acesso em: 6 maio 2014.

BECKER, K. ; TUMITAN, D. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios. Instituto de Informática UFRGS**, 2013. Disponível em: <http://www.inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf> Acesso em: 13 abr. 2014.

COPPIN, B. **Inteligência Artificial**. Tradução Jorge Duarte Pires Valério. 1. Ed. Rio de Janeiro: LTC, 2012.

DA COSTA, F. M. V. ; RALHA, J. C. L. ; RALHA C. G. Aprendizagem de Língua Assistida por Computador: Uma Abordagem Baseada em HPSG. **Revista Brasileira de Informática na Educação**, Brasília, v. 14, n.1, p. 20-21, jan./ abr. 2006.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 5ª ed. São Paulo: Atlas, 2010.

GONGORA, A. D. O que é Inteligência Artificial? .**egov.ufsc.br**, 2007. Disponível em: <http://www.egov.ufsc.br/portal/sites/default/files/anexos/6515-6514-1-PB.pdf> Acesso em 5 maio 2014.

LESSA, B. S. **Mineração de Opinião na Internet: Como Melhorar os Processos Internos da Empresa e Investir na Satisfação do Cidadão Através da Análise de Dados Coletados da Internet**. In: ConSerpro2012,Belém, 2012. Disponível em: <http://www.anaisdoconserpro.serpro.gov.br/modules/cadastro_de_trabalhos/trabalho.php?cod=224&ano=2012> Acesso em: 13 abr. 2014.

LOPES, T. Opsys! Mineração de Opiniões em Conteúdo Web. **opsys.com.br**, 2009. Disponível em: <http://www.opsys.com.br> Acesso em: 18 maio 2014.

LUGER, G. F. **Inteligência Artificial: Estruturas e Estratégias Para a Solução de Problemas Complexos**. Tradução Paulo Martins Engel. 4. Ed. Porto Alegre: Bookman, 2004.

MARTINS, C. A; MATSUBARA, E. T; MONARD, M. A**. PreTexT: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words**. In: IV International Workshop on Web and Text Intelligence, São Carlos, N.209, ago. 2003. Disponível em: <http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_209.pdf> Acesso em: 06 maio 2014.

MAZZOLA, Vitório B. **Engenharia de Software: Conceitos Básicos**. [S.l.:s.n.]. 199-?. 145 p. Apostila.

MELLO, S.C.B de; SÁ. M. G. de. Tecendo uma virtuosa "colcha de retalhos": a constituição e interpretação de um *corpus* linguístico num estudo sobre reflexividade e articulação empreendedora. **Revista de Administração Pública.** Rio de Janeiro. Jun. 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122006000300004&lng=en&nrm=iso&tlng=pt> Acesso em: 22 abr. 2014.

MODÉ, L. Cresce a competição entre empresas. **O Estado de São Paulo**, [São Paulo], 29 ago. 2010. Disponível em: <http://www.estadao.com.br/noticias/impresso,cresce-a-competicao-entre-empresas,601865,0.htm>. Acesso em: 25 mar. 2014.

NANNI, H.C.; CAÑETE, K. V. S**. A Importância das Redes Sociais como Vantagem Competitiva nos Negócios Corporativos**. In: XI Congresso Online de administração, 2014. Disponível em: <http://www.convibra.com.br/upload/paper/adm/adm_982.pdf> Acesso em: 26 mar. 2014.

NAVEGA, S. Inteligência Artificial, Educação de Crianças e o Cérebro Humano. **Revista de Estudos de Comunicações of the University of Santos,** São Paulo, n.72, fev. 2000. Disponível em: <http://www.intelliwise.com/reports/p4port.htm> Acesso em: 2 maio 2014.

OS NÚMEROS dos 10 anos de Facebook. **Adnews,** 2014. Disponível em: <http://www.adnews.com.br/internet/os-numeros-dos-10-anos-de-facebook> Acesso em: 26 mar. 2014.

PEROTTONI, R. et al. Sistemas de Informações: Um estudo Comparativo das Características Tradicionais às Atuais. **REAd** , Porto Alegre, n.3, jun. 2001. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/19461/000308562.pdf?sequence=1>Acesso em: 10 maio 2014.

PRESSMAN, R. S. **Engenharia de Software**. Tradução Ariovaldo Griesi, Mario Moro. 7. ed. Porto Alegre: AMHG, 2011.

ROSA, J. L. G. **Fundamentos da Inteligência Artificial**. 1. Ed. Rio de Janeiro: LTC,2011.

SARDINHA,T. B. **Linguística de corpus**. Barueri: Manole Ltda., 2004. Disponível em: <http://books.google.com.br/books?hl=pt-BR&lr=&id=i8uJXgeok48C&oi=fnd&pg=PR17&dq=corpus+o+que+%C3%A9&ots=R_70X_syPQ&sig=ncWkKMIgUh4NNJUQLX6FCKJ8HEc#v=onepage&q=corpus&f=false> Acesso em: 20 abr. 2014.

SILVA, B. C. D. da. et al. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações, **letras.etc.br,** 2007. Disponível em: <http://www.letras.etc.br/ebralc/NILCTR0710-DiasDaSilvaEtAl.pdf >Acesso em 2 maio 2014.

SOMMERVILLE, I. **Engenharia de Software**. Tradução Selma Shin Shimizu Melnikoff et al. 8. Ed. São Paulo: Pearson, 2007.

VON ZUBEN, F. J. Inteligência Artificial – Definição, Objetivos, Bibliografia, Histórico. **dca.fee.unicamp.br**, 2007. Disponível em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ea072_2s07/topico0_07.pdf> Acesso em: 10 maio 2014.

SMARTPHONE Samsung Galaxy S4 GT-I9505 desbloqueado. **Buscapé**, 2014. Disponível em: < http://www.buscape.com.br/avaliacao-positiva-sobre--smartphone-samsung-galaxy-s4-gt-i9505-desbloqueado.html>. Acesso em: 03 abr. 2014.

WEITZEL, L. et al. Somtools Uma Ferramenta Para Análise de Opinião e Sentimento no Ambiente Twitter. **infobrasil.inf.br,** 2013. Disponível em: <http://www.infobrasil.inf.br/userfiles/OK-SOMTOOLS-122373.pdf > Acesso em: 18 maio 2014.

ZILIO, D. Inteligência artificial e pensamento: redefinindo os parâmetros da questão primordial de Turing. **Ciência & Cognição,** Rio de Janeiro, n.1, mar. 2009. Disponível em: <http://pepsic.bvsalud.org/scielo.php?pid=S1806-58212009000100013&script=sci_arttext> Acesso em: 05 maio 2014.

APÊNDICE A – LISTA DE STOPWORDS

a, agora, ainda, alguem, ante, antes, ao, aos, apos, aquela, aquelas, aquele, aqueles, aquilo, as, ate, atraves, cada, coisa, com, como, contudo, da, daquele, daqueles, das, de, dela, delas, dele, deles, depois, dessa, dessas, desse, desses, desta, destas, deste, deste, destes, devera, deverao, deveria, deveriam, devia, deviam, disse, disso, disto, dito, diz, dizem, do, dos, e, é, ela, elas, ele, eles, em, enquanto, entre, era, essa, essas, esse, esses, esta, esta, estamos, estao, estas, estava, estavam, estavamos, este, estes, estou, eu, fazendo, fazer, feita, feitas, feito, feitos, foi, for, foram, fosse, fossem, ha, isso, isto, ja, la, lhe, lhes, lo, mas, me, mesma, mesmas, mesmo, mesmos, meu, meus, minha, minhas, na, nas, nem, nenhum, nessa, nessas, nesta, nestas, ninguem, no, nos, nossa, nossas, nosso, nossos, num, numa, nunca, o, os, ou, outra, outras, outro, outros, para, pela, pelas, pelo, pelos, pequena, pequenas, pequeno, pequenos, per, perante, pode, pude, podendo, poder, poderia, poderiam, podia, podiam, pois, por, porem, porque, primeiro, primeiros, propria, proprias, proprio, proprios, quais, qual, quando, quanto, quantos, que, quem, sao, se, seja, sejam, sem, sempre, sendo, sera, serao, seu, seus, si, sido, so, sob, sobre, sua, suas, talvez, tambem, tampouco, te, tem, tendo, tenha, ter, teu, teus, ti, tido, tinha, tinham, toda, todas, todavia, todo, todos, tu, tua, tuas, tudo, ultima, ultimas, ultimo, ultimos, um, uma, umas, uns, vendo, ver, vez, vindo, vir, vos.

Sistema de Análise de Sentimentos Baseada em Análise Superficial de Texto

Guilherme C. Marques, Patrick P. Silva, Elvio G. Silva, Henrique P. Martins

Universidade Sagrado Coração (USC)
Caixa Postal 511 – 17.011-160 – Bauru – SP – Brasil

Abstract. *The internet today is introduced as an great opinion source that permits prejudge the quality of a product. Due a huge amount of data, and before need for classify each found opinion, the human action in this task can be really slow. To remedy this problem, it is noticed the need to automate the classification process, what can be done via opinion mining techniques. So, the present study aimed to develop an algorithm for automatic opinion classification. To this, an opinion corpus was constructed and it served from basis to build a system that explore the surface text. The obtained results show that the system reached an 82% of precision in the opinion classification, demonstrating the potential of the method suggested.*

Resumo. *Hoje a internet se apresenta como uma grande fonte de opiniões que permitem pré-julgar a qualidade de um produto. Devido a enorme quantidade de dados, e diante da necessidade da classificação de cada opinião encontrada, a ação humana nesta tarefa pode acabar sendo muito lenta. Para sanar tal problema, nota-se a necessidade de automatizar o processo de classificação, o que pode ser feito por meio de técnicas de mineração de opinião. Assim, o presente trabalho teve como objetivo desenvolver um algoritmo para classificação automática de opiniões. Para isso um corpus de opiniões foi montado e serviu de base para construção de um sistema que explora os padrões de superfície de texto. Os resultados obtidos mostram que o sistema atingiu uma precisão de 82% na classificação de opiniões, mostrando o potencial do método sugerido.*

1. Introdução

Nos dias de hoje tornou-se comum e até previsível a disputa acirrada entre as grandes empresas em qualquer ramo no mercado. Segundo Modé (2010), "nunca foi tão forte na indústria brasileira a percepção de aumento da concorrência". Essa situação faz com que tais empresas tenham de tomar medidas para conhecer melhor o mercado, visando criar estratégias para melhorar as vendas. Sendo assim, é de vital importância para qualquer empresa saber a opinião do público sobre seus produtos já lançados, uma vez que deste modo, torna-se possível identificar quais pontos devem ser mudados e qual a melhor forma de lidar com as insatisfações.

Apesar de ser considerada uma ótima fonte de opiniões, devido a enorme quantidade de dados disponíveis, e diante da necessidade da classificação de cada opinião encontrada, a ação humana nesta tarefa, apesar de eficiente, pode acabar sendo muito lenta. Assim, para sanar tal problema, nota-se a necessidade de automatizar o processo de classificação, o que pode ser feito, por exemplo, por meio de técnicas de mineração de opinião (opinion mining), que são derivadas da mineração de dados. Segundo Becker e Tumitan (2013):

> A mineração de opinião é definida em como qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual.

Neste contexto, o presente trabalho teve como objetivo desenvolver, por meio de tal técnica, um algoritmo para classificação automática de opiniões. Tal sistema será baseado em um treinamento por meio de opiniões coletadas no site do "Buscapé".

2. Processamento de linguagem natural

O Processamento de Linguagem Natural (PLN) é o tratamento computacional dos aspectos da linguagem humana que leva em consideração formatos, estruturas e contextos. Segundo Rosa (2011, p.137) "[...] o Processamento de Línguas Naturais pode ser definido como a habilidade de um computador em processar a mesma linguagem que os humanos usam no dia a dia.", assim basicamente, pode-se dizer que o PLN visa fazer o computador se comunicar em linguagem humana, não necessariamente em todos os níveis de entendimento. (SILVA et al., 2007; DA COSTA; RALHA; RALHA, 2006; ROSA, 2011, grifo nosso).

O Processamento de Linguagem Natural tem uma grande variedade de aplicações possíveis e interessantes de serem abordadas. (SILVA et al., 2007). Uma delas é a utilização de tais sistemas em tarefas de análise de sentimentos, o que justifica a escolha do tema desta investigação.

Pode-se entender o estudo do PLN como uma "engenharia do conhecimento linguístico" e, deste modo, utilizar conceitos deste campo para o próprio desenvolvimento. (SILVA et al., 2007).

Assim, pode-se dividir o ato da concepção de um SPLN em três fases, sendo elas a "Fase Linguística", onde há um estudo mais aprofundado sobre a linguagem, montando-se um corpo de conhecimentos sobre ela, compreendendo todos os fenômenos linguísticos relevantes para o sistema. (SILVA et al., 2007, grifo nosso).

Após isso, há a "Fase Representacional", que é fase da construção conceitual do sistema, propondo sistemas formais de representação linguística e extralinguística, computacionalmente tratáveis (SILVA et al., 2007, grifo nosso).

E, por último, a "Fase Implementacional" que codifica todas as representações concebidas na fase anterior em linguagens de programação, além de fazer o planejamento global do sistema (SILVA et al., 2007, grifo nosso).

Tais fases visam principalmente, dentro da área de estudo do PLN, criar programas que facilitem a comunicação entre o usuário e o computador, utilizando-se de um conjunto de programas capazes de interpretar e gerar informações em mensagens linguisticamente construídas. (SILVA et al., 2007). Um dos recursos mais utilizados para a construção de SPLN são os corpora linguísticos.

2.1 Corpus

O conceito de corpus linguístico se baseia na coleta e exploração dos corpora, ou seja, conjuntos de dados linguísticos textuais que são coletados criteriosamente com o propósito de servir para a pesquisa de uma língua ou variedade linguística e, assim, dedicando-se a exploração da linguagem por meio de evidências empíricas, extraídas pelo computador. Pode-se entender então que um corpus linguístico é a representação de uma determinada realidade em determinado tempo, ou seja, é a representação de um contexto previamente definido para a pesquisa. (MELLO; SÁ, 2006; SARDINHA,2004).

2.2 Bag-Of-Words

A abordagem Bag-of-Words surgiu a partir da necessidade de organizar textos obtidos através da mineração de textos. A técnica visa facilitar o trabalho da aprendizagem de máquina, transformando através de um pré-processamento tais textos desestruturados em informações organizadas que poderão ser utilizadas pela maioria dos algoritmos com tal fim (MARTINS; MATSUBARA; MONARD, 2003).

3. Mineração de opinião

A mineração de opinião tem como objetivo a extração de opiniões e sentimentos de usuários sobre determinados temas ou produtos, contidas em sites especializados, para

posteriormente ocorrer a classificação de cada opinião, e descobrir o sentimento associado a cada uma delas. (AFONSO; GUEDES; MAGALHÃES, 2010; BARROS; LIMA; SILVA, 2012).

4. Análise de Sentimentos

A Análise de Sentimentos surge praticamente como sinônimo de mineração de opiniões, e tem como objetivo identificar o sentimento expresso em textos opinativos, textos esses que são subjetivos, uma vez que não apresentam fatos concretos como os objetivos. (BARROS; LIMA; SILVA, 2012).

Ela pode ser realizada em três níveis que diferenciam profundidades dentro do texto. O nível de documento, onde observa o sentimento expresso no texto como um todo. Em seguida o nível de sentença classifica a polaridade de cada sentença dentro do texto. E em nível mais profundo, o nível de característica que indica a polaridade de cada atributo do objeto analisado, conseguindo uma visão refinada de cada opinião. (BARROS; LIMA; SILVA, 2012).

Para os comentários em formato de prós e contras, são necessárias técnicas de aprendizado de máquina, empregada a um corpus com as opiniões já classificadas, conforme a divisão encontrada na fonte. (BARROS; LIMA; SILVA, 2012).

A classificação de sentimentos é a principal etapa dentro de uma análise de sentimentos, pois será ela que identificará de fato a polaridade do texto. (BARROS; LIMA; SILVA, 2012).

De acordo com os autores supracitados, assim como na extração de características, a abordagem de aprendizado de máquina também aparece na classificação de sentimentos, necessitando de um corpus de treinamento precisamente dividido e etiquetado, sendo necessário categorizar a polaridade de expressões de menor frequência, o que a torna ao mesmo tempo mais precisa e também muito mais custosa.

Todo esse processo de análise de sentimentos pode ser realizado no contexto de web mining, permitindo a utilização de sites online para identificação e monitoramento de polaridade em mensagens compartilhadas que carregam emoções expressas pelos usuários. Assim, considerando o foco deste trabalho e o uso potencial do processo de mineração de opinião, o site Buscapé, que pode ser considerado uma fonte de opiniões para construção de um corpus voltado ao processo de desenvolvimento de uma ferramenta de análise de sentimentos.

5. Metodologia

Este projeto é uma pesquisa exploratória, pois estudou e implementou uma abordagem que possibilitou o desenvolvimento de um protótipo de sistema de análise de sentimentos para a língua portuguesa. Basicamente, para cumprimento de tal objetivo foram necessários três passos: a definição da arquitetura, funcionamento do sistema e o processo de formação de um corpus linguístico (GIL, 2010)

Tendo em vista a recuperação e manipulação das opiniões citadas anteriormente, o presente trabalho foi dividido em cinco processos conforme a Figura 1 ilustra.

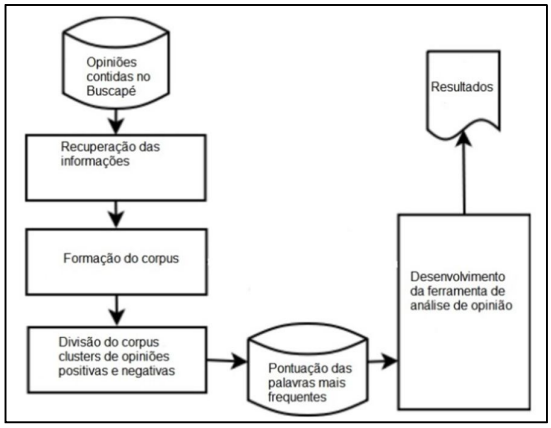


Figura 1 – Arquitetura do projeto

Assim como ilustrado na Figura 1, para a elaboração do protótipo do sistema de análise de sentimento, em primeiro lugar houve a recuperação das opiniões contidas dentro do site do "Buscapé", tal processo ocorreu maneira automatizada, o que possibilitou a formação do corpus linguístico, que foi o resultado da união de todas as opiniões obtidas a partir da recuperação. Esse corpus foi dividido em dois clusters com a abordagem de Bag-Of-Words, um contendo as opiniões positivas e outro contendo as negativas. Posteriormente, tais clusters foram submetidos à ferramenta de análise de corpus, o "Corpógrafo", que indicou a frequência de unigramas[5], bigramas e trigramas mais significativos, que puderam, assim, ser pontuados (recebendo um peso), visando criar um ranking de palavras indicativas de opinião que foram utilizadas no processo de análise de sentimentos. Para esta investigação foi considerado que a palavra/expressão que possuiu a maior frequência dentro de um cluster positivo ou negativo, receberia o peso máximo (1 para positivo e -1 para negativo); e as demais receberam pesos proporcionais à sua frequência no corpus. A utilização de pesos entre -1 e 1 foi escolhida por ser algo comum em sistemas de processamento de língua natural que utilizam frequências relativas no processo de ponderação de textos. As expressões repetidas entre ambos tiveram seus pesos somados. Tal pontuação de expressões serviu como base para o analisador de sentimentos, que comparou as palavras e expressões da opinião recebida com as de sua base, atribuindo o peso pré-estabelecido no processo anterior, tornando, assim, possível a soma de todos esses pesos, possibilitando através do resultado final, determinar se tal opinião é positiva ou negativa. Ao fim deste processo, opiniões cujas pontuações (soma dos pesos recebidos) estejam acima de zero, provavelmente indicam um caráter positivo, já aquelas abaixo de zero tem um caráter negativo.

Para haver a criação do corpus, foram utilizadas cerca de cinco mil opiniões Entretanto, não foi adotado nenhum critério de balanceamento no número de opiniões positivas e negativas. Captadas através de uma extensão do navegador "google chrome" chamada "Web Scraper", que tem o papel de capturar de forma automática o conteúdo de determinadas divisões dentro de uma página indicada, retornando um arquivo tipo CSV para cada produto em específico. Tais arquivos foram posteriormente reunidos em um único arquivo TXT para opiniões positivas e para as opiniões negativas.

Após a união de todas as opiniões em um único arquivo de texto para cada polaridade, tais arquivos foram submetidos a um processo de pré-processamento para diminuir a probabilidade de conflitos referentes à acentuação, caracteres especiais e entre letras maiúsculas e minúsculas. Além disso, também foram retiradas as stopwords, que são palavras que não adicionam sentido nenhum dentro qualquer possível contexto. Tal processo se deu de

[5] Os unigramas são palavras únicas, já os bigramas e trigramas são, respectivamente, sequências de duas ou três palavras.

forma automatizada por meio de uma aplicação criada especificamente para este fim, utilizando-se a linguagem de programação PHP.

Destaca-se que optou-se por uma filtragem de termos realizada de modo manual, para que houvesse maior controle dos dados que seriam utilizados posteriormente, priorizando a qualidade dos dados. Para a lista de unigramas, por se tratar de palavras únicas, a ferramenta acabou por listar todas as palavras contidas no corpus, deste modo definiu-se que para o presente projeto seriam utilizadas em torno das cinquenta mais frequentes, com exceção das palavras que denotam marcas como Samsung, Apple ou Sony, e também palavras relacionadas a nomes de modelos de smartphones como Iphone, Moto X ou Nexus.

Com relação aos bigramas e, principalmente aos trigramas, este processo se deu de maneira a serem escolhidas as expressões que fariam sentido e que representassem a real ideia de seu corpus como, por exemplo, ao se analisar os bigramas positivos foram desconsideradas expressões como “não gostei” ou “deixa desejar”. Além de haver o mesmo critério de descartar marcas e nomes de modelos utilizado nos unigramas, sendo aproveitadas todas as expressões consideradas relevantes.

Assim, ao final de todo este processo, foram consideradas para a aplicação um total de 98 unigramas, sendo 50 positivos e 48 negativos, 70 bigramas, 48 positivos e 22 negativos e 32 trigramas, sendo 29 positivos e 3 negativos.

Finalizada a etapa de análise das expressões que seriam utilizadas no processo de pontuação pela aplicação, foram estabelecidos os pesos que seriam atribuídos a cada expressão. Primeiramente, como citado anteriormente, definiu-se que as expressões teriam peso entre 1 e -1. Para isso, todas as expressões mais frequentes tiveram atribuídos a elas peso 1 positivo, se fosse um grupo de expressões positivas e peso -1 se fossem dos grupos negativos. Deste modo, para as expressões subsequentes, seu peso foi definido pela divisão de sua frequência, pela frequência da expressão com peso 1, por exemplo, para os bigramas positivos, a expressão mais frequente foi “muito bom”, ocorrendo 697 vezes dentro do corpus, e a expressão “ótimo produto” ficou em quinto lugar com 174 ocorrências. Desta maneira, para se definir o peso da expressão “ótimo produto” dividiu-se 174 por 697, o que lhe gerou de peso final 0,24. Este procedimento foi aplicado para todas as outras expressões obtidas nas etapas anteriores, gerando o peso de cada uma delas.

6. Arquitetura e funcionamento do sistema

Para possibilitar o desenvolvimento do software, todas as expressões selecionadas e pontuadas, foram inseridas dentro de um banco de dados MySQL, tal escolha para o sistema de gerenciamento de banco de dados se deve ao fato de ser um sistema gratuito, além de ter grande compatibilidade com a linguagem de programação adotada para o software, o PHP.

O software é constituído em quatro fases simples, a entrada de dados via interface, como ilustrado na Figura 2, o pré-processamento do texto de entrada, a análise e o retorno, mostrados na Figura 3.

Figura 2 – Interface do Software

Como pode ser observado a partir da Figura 2, a interface do programa é muito simples, constituída de uma área para a inserção do texto e dois botões, um visando a classificação, e outro para limpar a área de texto, além do conteúdo textual indicando a função do programa. Nesta área o usuário deverá inserir sua opinião a fim de classificá-la.

Após a ativação do botão “classificar”, o programa entra em um estado de processamento que não está visível para o usuário.

Nesta fase, primeiramente o texto inserido passará por um pré-processamento, onde a opinião inserida terá um tratamento idêntico ao dado para o corpus anteriormente, exatamente com o propósito de deixar os textos que serão comparados em formatos semelhantes, sem acentos, caracteres especiais e sem as stopwords, minimizando, assim, as possibilidades de conflito.

Já com o texto pré-processado, esta opinião será disposta em um vetor, como uma Bag-of-words, que será comparado aos vetores recuperados do banco de dados com as expressões do corpus na fase de análise. Cabe destacar que optou-se inicialmente por utilizar de forma concomitante, tanto os unigramas quanto bigramas e trigramas para a classificação da opinião. A hipótese era de que os três de forma conjunta seriam capazes de classificar satisfatoriamente, uma opinião.

Para a pontuação da opinião, o vetor formado será percorrido, e comparado aos vetores recuperados das tabelas inseridas no banco de dados, primeiramente aos trigramas, bigramas e unigramas (nesta ordem) positivos, e depois pelo mesmo processo para os negativos. A comparação consiste em confrontar as expressões, sendo que se ela estiver contida no vetor recuperado do banco, a opinião receberá a pontuação dada à expressão anteriormente, somando-se assim à que já havia sido atribuída.

Ao final, somam-se os resultados obtidos a partir das expressões positivas e negativas, e a que obtiver uma pontuação maior, definirá a polaridade da opinião, ou seja, se ela é positiva ou negativa.

Após o fim da fase de análise e com a opinião classificada, o programa mostrará o resultado obtido ao usuário por meio de um alerta, como na Figura 3.

Figura 3 – Retorno do Software

Para a avaliação do sistema, as etapas de interface e retorno foram retiradas do processamento, para possibilitar que o processo fosse feito de forma automatizada. Tal avaliação será descrita de maneira mais aprofundada na próxima seção.

7. Resultados

Ao fim do desenvolvimento, a aplicação foi submetida a testes de precisão que visaram verificar a sua eficiência na tarefa de classificação. Os testes também foram definidos de forma a verificar se a hipótese inicial de utilizar de forma conjunta, tanto os unigramas quanto bigramas e trigramas para a classificação da opinião traria, de fato, os melhores resultados, ou se haveria alguma outra combinação que pudesse melhorar a precisão do sistema.

Para esta verificação, um corpus de teste foi formado com opiniões disponíveis em vários sites e-commerce na internet, tais como “americanas.com.br”, “shoptime.com.br”, dentre outros. Visando uma maior confiabilidade, não foram utilizadas opiniões contidas no site do “Buscapé”, uma vez que o mesmo serviu como repositório do corpus de treinamento da aplicação.

No total o corpus de teste foi formado por 100 opiniões, sendo 50 positivas e 50 negativas. O sistema foi testado em diferentes situações, avaliando-se sempre a porcentagem de acertos (precisão) em cada uma das situações propostas.

No total a aplicação foi testada de oito maneiras diferentes, sendo elas: Teste 1: utilizando todo o repositório formado ao logo do processo, portanto, unigramas, bigramas e trigramas positivos e negativos; Teste 2: utilizando a tabela de unigramas negativos modificada e sem a palavra “não” inclusa, uma vez que observou-se que houve grande diferença de frequência entre a palavra “não” e as demais, sendo importante avaliar a sua influência no resultado; Teste 3: não foram retiradas as “stopwords” na etapa de pré-processamento da opinião; Teste 4: a opinião foi confrontada apenas com as tabelas de unigramas; Teste 5:apenas os bigramas foram comparados à opinião; Teste 6: foram unidos os unigramas aos bigramas; Teste 7: uniu os unigramas e os trigramas; Teste 8: que testou a união dos bigramas com os trigramas.

A tabela 2 a seguir mostra resumidamente todos os testes realizados para o sistema, dando uma visão geral dos resultados obtidos.

	Teste 1	Teste 2	Teste 3	Teste 4	Teste 5	Teste 6	Teste 7	Teste 8
Total de acertos:	79	60	80	82	30	80	82	30
Total de erros:	20	38	19	17	5	19	17	6
Não classificadas:	1	2	1	1	65	1	1	64
Opiniões negativas classificadas como positivas:	8	33	8	4	0	7	4	0
Opiniões positivas classificadas como negativas:	12	5	11	13	5	12	13	6
Precisão:	79,00%	60,00%	80,00%	82,00%	30,00%	80,00%	82,00%	30,00%

Tabela 2 – Resumo dos resultados

Ao final de todos os testes, foi possível notar a importância dos unigramas dentro do sistema que utiliza o método de classificação proposto nesta investigação, uma vez que eles potencializam a taxa de precisão quando utilizados sozinhos e fazendo-a diminuir quando não utilizados, como pode ser visto a partir da Figura 16. Tal ocorrência se deve ao fato unigramas serem mais facilmente reconhecidos dentro de um texto, uma vez que são palavras únicas e no presente projeto estarem em maior número do que os outros tipos de expressões.

8. Considerações finais

Os resultados obtidos a partir do presente trabalho contribuem para pesquisas no campo do processamento de linguagem natural em língua portuguesa, uma vez que o método proposto obteve uma taxa de precisão de 82% na classificação das opiniões submetidas ao sistema, o que pode ser considerado como uma abordagem com um bom potencial.

Porém o presente método também apresentou limitações, como as indicadas durante a discussão dos resultados, como a deficiência do sistema em classificar as opiniões corretamente sem a inclusão dos unigramas. Isso decorre, provavelmente, do fato de haverem poucas opções de expressões dentre bigramas e trigramas. Como possibilidade de extensão desta pesquisa, sugere-se um foco maior na coleta de mais expressões deste tipo, o que pode gerar melhores resultados.

Outra limitação encontrada refere-se ao processo de coleta do repositório de opiniões, em que o número de opiniões positivas encontradas foi muito maior do que o número de negativas, o que afetou a diversidade e número de expressões negativas utilizadas no sistema. Como trabalhos futuros com a mesma temática sugere-se a utilização de outro repositório de opiniões, mais consistente, balanceado e com número maior de opiniões negativas em seu escopo. Também poderá ser realizada por especialistas uma coleta mais seletiva das opiniões utilizadas para o corpus, que pode melhorar a qualidade do repositório.

Deste modo, conclui-se que o método proposto e implementado por meio de um software, apesar de simples, tem um grande potencial, já que se alcançou bons índices de precisão ao classificar opiniões, sendo pouco custoso já que trabalha apenas no nível superficial do texto. Deste modo, o presente trabalho contribui para pesquisas ligadas ao processamento de linguagem natural, principalmente em língua portuguesa, uma vez que há poucos trabalhos específicos nesta língua. Porém ainda há muito a ser melhorado e novas propostas podem contribuir ainda mais para o crescimento de pesquisas deste tipo.

Referências

AFONSO, D.; GUEDES, R.; MAGALHÃES, L. H. de. Mineração de Opiniões de Usuários na Busca de Conhecimento, Revista das Faculdades Integradas Vianna Júnior, Juiz de Fora, out. 2010. Disponível em: <http://www.viannajr.edu.br/files/uploads/20131001_141137.pdf> Acesso em: 08 maio 2014.

BARROS, F.; LIMA, D.; SILVA, N. R. SAPair: Um Processo de Análise de Sentimento no Nível de Característica, labic.icmc.usp.br, 2012. Disponível em: <http://www.labic.icmc.usp.br/wti2012/artigos/105283.pdf>Acesso em: 6 maio 2014.

BECKER, K. ; TUMITAN, D. Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios. Instituto de Informática UFRGS, 2013. Disponível em: <http://www.inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf> Acesso em: 13 abr. 2014.

DA COSTA, F. M. V. ; RALHA, J. C. L. ; RALHA C. G. Aprendizagem de Língua Assistida por Computador: Uma Abordagem Baseada em HPSG. Revista Brasileira de Informática na Educação, Brasília, v. 14, n.1, p. 20-21, jan./ abr. 2006.

GIL, A. C. Como Elaborar Projetos de Pesquisa. 5ª ed. São Paulo: Atlas, 2010.

MARTINS, C. A; MATSUBARA, E. T; MONARD, M. A. PreTexT: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. In: IV International Workshop on Web and Text Intelligence, São Carlos, N.209, ago. 2003. Disponível em: <http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_209.pdf > Acesso em: 06 maio 2014.

MELLO, S.C.B de; SÁ. M. G. de. Tecendo uma virtuosa "colcha de retalhos": a constituição e interpretação de um corpus linguístico num estudo sobre reflexividade e articulação empreendedora. Revista de Administração Pública. Rio de Janeiro. Jun. 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122006000300004&lng=en&nrm=iso&tlng=pt> Acesso em: 22 abr. 2014.

MODÉ, L. Cresce a competição entre empresas. O Estado de São Paulo, [São Paulo], 29 ago. 2010. Disponível em: <http://www.estadao.com.br/noticias/impresso,cresce-a-competicao-entre-empresas,601865,0.htm>. Acesso em: 25 mar. 2014.

ROSA, J. L. G. Fundamentos da Inteligência Artificial. 1. Ed. Rio de Janeiro: LTC,2011.

SARDINHA,T. B. Linguística de corpus. Barueri: Manole Ltda., 2004. Disponível em: <http://books.google.com.br/books?hl=pt-BR&lr=&id=i8uJXgeok48C&oi=fnd&pg=PR17&dq=corpus+o+que+%C3%A9&ots=R_70X_syPQ&sig=ncWkKMIgUh4NNJUQLX6FCKJ8HEc#v=onepage&q=corpus&f=false> Acesso em: 20 abr. 2014.

SILVA, B. C. D. da. et al. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações, letras.etc.br, 2007. Disponível em: <http://www.letras.etc.br/ebralc/NILCTR0710-DiasDaSilvaEtAl.pdf >Acesso em 2 maio 2014.