UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA – DEPARTAMENTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM DESENVOLVIMENTO DE SISTEMAS PARA WEB

FRANK WILLIAN CARDOSO DE OLIVEIRA

ANÁLISE DE SENTIMENTOS DE COMENTÁRIOS EM PORTUGUÊS UTILIZANDO *SENTIWORDNET*

MARINGÁ
2013

UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA – DEPARTAMENTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM DESENVOLVIMENTO DE SISTEMAS PARA WEB

FRANK WILLIAN CARDOSO DE OLIVEIRA

ANÁLISE DE SENTIMENTOS DE COMENTÁRIOS EM PORTUGUÊS UTILIZANDO *SENTIWORDNET*

Trabalho submetido à Universidade Estadual de Maringá como requisito para obtenção do título de Especialista em Desenvolvimento de Sistemas para *Web*. Orientador: Prof. Dr. Sérgio Roberto P. da Silva.

MARINGÁ
2013

UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA – DEPARTAMENTO DE INFORMÁTICA
ESPECIALIZAÇÃO EM DESENVOLVIMENTO DE SISTEMAS PARA WEB

FRANK WILLIAN CARDOSO DE OLIVEIRA

ANÁLISE DE SENTIMENTOS DE COMENTÁRIOS EM PORTUGUÊS UTILIZANDO *SENTIWORDNET*

Aprovado em ____/____/____

BANCA EXAMINADORA

_____
Prof.ª Dr.ª Valéria Delisandra Feltrim
Universidade Estadual de Maringá

_____
Prof. Dr. Wagner Igarashi
Universidade Estadual de Maringá

_____
Prof. Dr. Sérgio Roberto P. da Silva (Orientador)
Universidade Estadual de Maringá

MARINGÁ

2013

RESUMO

A constante utilização da *Web* e, consequentemente, dos serviços por ela oferecidos diminuiu distâncias e facilitou a comunicação entre as pessoas. Este fator tem despertado e contribuído para o surgimento de novas áreas de pesquisa que se utilizam da *Web* como base para a coleta de dados. O elemento motivador é o fato de que as pessoas se sentem mais livres, e com mais oportunidades, para expor suas opiniões em meios eletrônicos e, geralmente, essas mensagens ficam disponíveis para visualização por inúmeras pessoas. Juntamente a este cenário surgiu à análise de sentimentos, a qual tem como objetivo principal classificar comentários geralmente referentes a uma entidade, a qual pode ser um produto, serviço ou pessoa. Neste contexto, foi proposto um protótipo para classificar comentários em português, independente do domínio de negócio, tendo como base de dados os comentários do *Twitter*®. Para realizar a classificação foi utilizada a *SentiWordNet*®, uma base de abreviaturas e gírias, e outros recursos para o pré-processamento dos textos. Concluiu-se que alguns fatores influenciaram em diferentes graus no resultado da taxa de acertos como, por exemplo, a necessidade de tradução dos documentos, por não haver bases em português; a ambiguidade nos comentários; as gírias as abreviaturas e/ou termos informais utilizados constantemente.

PALAVRAS-CHAVE: Análise de Sentimentos; *SentiWordNet*®.

ABSTRACT

The constant use of the Web and, consequently, the services it offered reduced distances and facilitated communication between people. This factor has raised and contributed to the emergence of new research areas that use the Web as a basis for data collection. The motivating factor is the fact that people feel freer, and more opportunities to express their opinions in electronic media and, generally, these messages are available for viewing by numerous people. Along with this scenario came to sentiment analysis, which aims to classify comments primary usually referring to an entity, which can be a product, service or person. In this context, we propose a prototype to sort comments in Portuguese, domain independent business, based on data from Twitter® comments. To perform the classification was used SentiWordNet®, a database of abbreviations and slang, and other resources for pre-processing of texts. It was concluded that some factors influencing the results in varying degrees of accuracy rate, for example, the need to translate documents, because there is no basis in Portuguese; ambiguity in the comments; slang abbreviations and/or terms used informal constantly.

KEYWORDS: Sentiment Analysis; SentiWordNet®.

LISTA DE FIGURAS

Figura 1 – Etapas comum de um analisador de sentimentos. 16
Figura 2 - Resultado apresentado pelo Sentiment140® 17
Figura 3 - Apresentação de resultado proposto por (PANG e LEE, 2008). 18
Figura 4 - Resultado de pesquisa no website SentiWordNet®. 20
Figura 5 - Arquitetura do projeto desenvolvido. 22
Figura 6 - Comentário no Twitter®. 28
Figura 7 - Resultado pós Processamento 28
Figura 8 - Comentários com as classificações. 29
Figura 9 - Comentários retirado do Twitter® 35

LISTA DE TABELAS

Tabela 1 - Estrutura da SentiWordNet©. 20
Tabela 2 - Alguns parâmetros aceito pela API do Twitter®. 24
Tabela 3 - Estrutura do dicionário de abreviaturas e siglas. 26
Tabela 4 - Termos pesquisados. 32
Tabela 5 - Classificações dos comentários retirados do Twitter®. 33
Tabela 6 – Matriz de Confusão – Comentários Twitter®. 34
Tabela 7 - Exemplo de opiniões retiradas do Twitter®. 35
Tabela 8- Classificação de comentários - Youtube®. 37
Tabela 9 – Matriz de Confusão - Youtube®. 37
Tabela 10 - Classificação de comentários - Mercado Livre®. 38
Tabela 11 – Matriz de Confusão - Mercado Livre® 38

SUMÁRIO

1 – INTRODUÇÃO.....7

2 – ANÁLISE DE SENTIMENTOS E/OU OPINIÕES..... 12

2.1 – As Fontes de Dados.....13

2.2 – Dificuldades Encontradas e Pontos a Ser Considerados para a Análise de Sentimentos..... 14

2.3 – Aplicações com Análise de Sentimentos.....15

2.4 – Etapas da Análise de Sentimentos.....16

2.5 – Técnicas Baseadas em Recursos Léxicos.....18

2.6 – A *SentiWordNet*© 19

3 – DESCRIÇÃO DO SISTEMA DESENVOLVIDO.....21

3.1 – Arquitetura Geral do Sistema.....22

3.2 – O Módulo de Busca.....23

3.3 – O Módulo de Pré-Processamento.....25

3.3.1 – Dicionário de Abreviatura e Gírias.....25

3.3.2 – Tradução..... 26

3.3.3 – O *Stemming*..... 27

3.4 - A Classificação dos Comentários..... 27

3.5 – O Módulo Analisador.....28

3.6 – Tratamento das Palavras não Encontradas..... 30

4 –Avaliação do Protótipo Proposto.....31

4.1 – Metodologia de Avaliação..... 31

4.2 – Os Testes do Protótipo..... 33

5 – CONSIDERAÇÕES FINAIS..... 40

REFERÊNCIAS..... 42

1 – INTRODUÇÃO

O advento das mídias sociais por meio da *Web* tornou fácil a interação entre as pessoas, a publicação de conteúdos e a exposição de opiniões, contribuindo para a *Web* se tornar um grande repositório de dados, principalmente em estruturas no formato de textos (PANG e LEE, 2008). Até o final de 2011 o serviço de *microblog Twitter*® registrou, somente no Brasil, um total de 33,3 milhões de usuários e outro meio de comunicação social, o *Facebook*®, apresentou um crescimento de 298%, chegando a 35 milhões de usuários (G1, 2012). O comercio eletrônico também apresentou grande um crescimento, registrando um aumento de 26%, chegando a um montante negociado no valor de 18,7 bilhões de reais no ano de 2011 em relação a 2010, tendo sido realizado um total de 53,7 milhões de pedidos via *Web* (SOLUCIONA, 2012).

A quase onipresença da *Web* na vida das pessoas hoje em dia e, consequentemente, dos serviços por ela oferecidos, tem contribuído para que seus usuários compartilhem constantemente inúmeras mensagens, relatando suas experiências referentes às estadias em determinados hotéis; ou comentando nas redes sociais sobre o carro, o micro-ondas, o refrigerador ou o celular novo que acabou de adquirir, etc. Devido a esses fatores, a carga de dados da *Web* cresce a todo o momento de forma estrondosa. Esse crescimento gera um montante de informações que pode ser útil para que as organizações detentoras dos produtos, ou serviços, possam saber, por exemplo, se seus clientes estão, ou não, satisfeitos. Uma das vantagens em utilizar a *Web* para coletar esse tipo de informação é o fato de as pessoas se sentirem livres para expor suas ideias e experiências quando a fazem na *Web*.

A *Web* tem ampliado as possibilidades de obtenção de informação no nosso dia-a-dia. Por exemplo, quando uma pessoa está interessada em comprar algum produto, ou serviço, ela normalmente procura informação com os amigos, ou pessoas, que ela sabe que possuem, ou utilizaram, o produto ou o serviço. As opiniões são os principais influenciadores do comportamento humano LIU, (2007). Assim, é de grande interesse saber o que as pessoas estão pensando e, consequentemente, falando a respeito de produtos e serviços. Quando essa compra se realiza pela *Web*, caso em que nem sempre

se conhece fisicamente o produto e/ou a loja que o está vendendo, é comum as pessoas buscarem opiniões nas revisões fornecidas por outras pessoas (as quais muitas vezes ela nem conhece), no *website* da própria loja ou nas redes sociais.

Porém, buscar e filtrar dados produzidos na *Web* para extrair informações relevantes para ajudar uma organização, e/ou usuários, em tomadas de decisão não é algo simples de ser executado de forma manual. Primeiramente porque muitos conteúdos da *Web* se encontram desestruturados, e segundo pela velocidade com que a quantidade de dados na *Web* cresce. Para varrer milhares de *post* do *Twitter*® constantemente, por exemplo, de forma automática e verificar se esses são favoráveis ou não a imagem de uma empresa, produto e/ou serviço, é necessário combinar várias tecnologias e técnicas, e desenvolver um processo automático.

Desde o surgimento da *Web* 2.0, surgiu também uma nova área de pesquisa denominada mineração de opinião, ou análise de sentimentos, tendo como objetivo desenvolver sistemas para analisar as opiniões, avaliações, atitudes e emoções das pessoas em relação a certa entidade (LIU, 2007). Basicamente, uma análise de sentimentos classifica o comentário sob análise como positivo, negativo ou neutro.

O primeiro desafio para o desenvolvimento de um sistema de análise de sentimentos é a escolha da fonte de dados da qual as informações serão extraídas, o que é de grande importância para que as informações apresentadas ao usuário sejam confiáveis. A confiabilidade é um dos fatores preponderantes nesse tipo de sistema, pois muitas vezes as informações que chegam aos usuários podem apresentar desvios tendenciosos. Por exemplo, normalmente as empresas publicam em redes sociais, *websites* e outros meios de comunicação, conteúdos relativos a propaganda de seus produtos. Supondo que um usuário pesquise sobre esse produto, essa publicação pode gerar um falso sentimento positivo.

Outros problemas comumente encontrados no processamento de notícias divulgadas nas redes sociais, são o uso de ironias nos comentários e as distribuições de *spams*, os quais, respectivamente, tornam difícil a identificação automática da semântica da mensagem e são falsos por sua natureza.

Uma das questões a ser considerada em sistemas que realizam a análise de sentimentos é como será realizada a classificação das informações. De acordo com LIU, (2007) uma das abordagens mais utilizadas são as técnicas baseadas em aprendizagem supervisionada. Como exemplo pode-se citar o algoritmo de *Support Vector Machine*

(*SVM*) e o classificador *Naive Bayes*. Liu (2007) utilizou o algoritmo de aprendizagem de máquinas *SVM* para desenvolver um projeto para classificar comentários de filmes em duas classes, positivos e negativos. Após avaliações o projeto obteve, segundo o autor, um bom desempenho com taxas de acerto que variam entre 72,8% a 78,7%. O *Sentiment140*® é outro projeto que realiza análise de sentimentos, fazendo uso de um classificador de *Maximum Entropy* (SENTIMENT140, 2012). Ele está disponível para a língua inglesa e espanhola e possui também uma opção para as empresas monitorarem suas marcas e produtos. Ele utiliza como fonte de dados exclusivamente o *Twitter*® (https://twitter.com). Como ferramenta para a língua portuguesa pode-se citar o O*pSys*®, o qual em sua primeira versão tem o foco voltado para as empresas negociadoras da Bovespa, tendo como fonte de dados textos de *RSS* (LOPES, 2009). Segundo os autores, essa ferramenta trabalha com um algoritmo parametrizável, o que torna mais simples a mudança de domínio.

Outra abordagem para análise de sentimentos disponível na literatura é o emprego de recursos léxicos (OHANA, TIERNEY, 2009), geralmente utilizando-se de uma base composta por termos opinativos que são previamente classificados. O *SentiWordNet*© é um exemplo desses recursos (BACCIANELLA, ESULI, SEBASTIANI, 2010), o qual é um dicionário construído utilizando-se o método de aprendizagem semi-supervisionado, extraindo os termos da *WordNet*© e atribuindo a cada *synset* duas notas: uma positiva e uma negativa. Atualmente ele está disponível somente para língua inglesa. Um dos pontos positivo em utilizar esse tipo de abordagem é o fato de que o sistema será livre de domínio, pois a análise é realizada em cima dos termos que compõe o texto a ser analisado, os quais, consequentemente, devem se encontrar na base do *SentiWordNet*©. Por outro lado, muitos domínios de negócios possuem características particulares que influenciam nos resultados como, por exemplo, o termo “gelado” que é positivo quando se fala em cerveja e negativo quando se trata de pizza. Nesse contexto, a análise léxica é prejudicada, pois o domínio de negócio não é levado em consideração.

As aplicações que utilizam algoritmos de aprendizagem de máquina possuem uma taxa de acerto maior, quando comparado a outras técnicas como, por exemplo, recursos léxicos. Esse fato é devido ao treinamento realizado no classificador antes de seu uso, o qual geralmente foca um domínio especifico como, por exemplo, o de alimentos, locais, pessoas, etc.

Do ponto de vista de análise de sentimentos em redes sociais, Souza (2011) faz uma análise de comentários do *Twitter*® da língua inglesa, tendo conseguido uma taxa de acerto de aproximadamente 54%. Já o trabalho apresentado por DENECKE, (2008), trata da utilização do *SentiWordNet*© para qualquer linguagem por meio de tradução para a língua inglesa. O método de classificação utilizado foi dividido nas seguintes etapas: classificação da linguagem, tradução do documento, preparação do texto e a análise propriamente dita. O autor realizou vários testes de exatidão no analisador trabalhando com comentários de filmes. Os testes foram realizados para a língua inglesa e a alemã, tendo obtido respectivamente os valores 51% a 62% e 59% a 66% de acertos. O autor ainda concluiu que a tradução dos documentos influência nos resultados, o que acontece muitas vezes é uma mudança no sentido do texto traduzido em relação ao texto original.

Após a realização de buscas e estudos da literatura, foi constatado que poucos trabalhos foram desenvolvidos para análise de sentimentos em textos da língua portuguesa. Desse modo, tendo como base o trabalho de DENECKE, (2008), este trabalho tem por objetivo geral o estudo da área de análise de sentimentos e o desenvolvimento de um protótipo de um sistema para realização de análise de comentários na língua portuguesa, utilizando-se do *SentiWordNet*© para a avaliação da polaridade dos termos presentes nos *tweets*. Para tal deverá ser empregado um esquema de tradução da língua portuguesa para a língua inglesa antes de se submeter o termo para análise pelo *SentiWordNet*©.

Para alcançar este objetivo foi utilizada a seguintes metodologia:

- Investigação das tecnologias envolvidas na análise de sentimentos e estudo teórico do assunto.
- Análise e definição de quais as tecnologias e ferramentas a serem utilizadas para desenvolver o protótipo.
- Desenvolvimento de um protótipo.
- Montagem de uma base de comentários para os testes, na qual cada comentário será classificado manualmente.
- Realização de testes utilizando a base criada na etapa anterior a fim de verificar a eficiência do sistema.

A fonte de dados escolhida para realizar os testes de desempenho foi o *Twitter*®, pelo fato de que muitas pessoas utilizam constantemente esse meio de comunicação para expor suas opiniões referentes às experiências vivenciadas diariamente.

Esta monografia está dividia em quatro capítulos seguidos de uma conclusão, sendo o primeiro a introdução. O Capítulo 2 traz uma contextualização da área de análise de sentimentos. O Capitulo 3 mostra o protótipo proposto neste trabalho, sendo apresentada a arquitetura e explicado cada módulo que a compõe. O Capítulo 4 apresenta os testes de avaliação e os resultados que foram realizados com o protótipo. Por último, apresentamos nossa considerações finais e algumas sugestões de trabalhos futuros.

2 – ANÁLISE DE SENTIMENTOS E/OU OPINIÕES

Quando uma pessoa tem interesse em algum produto e/ou serviço, é comum que ela entre em *websites* de *reviews*, *blogs*, redes sociais, entre outras fontes de informações na *Web*, buscando verificar as opiniões de outras pessoas em relação ao produto de interesse para, a partir dessas opiniões, fazer uma análise e decidir se deve ou não comprar o produto.

Esta busca e análise podem ser realizadas de forma manual se o número de opiniões for pequeno. No entanto, ela se torna trabalhosa à medida que a quantidade de informação vai aumentando. Para tornar a tarefa mais fácil surgiu uma nova área de pesquisa denominada análise de sentimentos e/ou opiniões, a qual é definida por LIU, (2007) como a área da computação que estuda as opiniões, avaliações e atitudes das pessoas expressas em textos, sempre direcionada a uma entidade, a qual pode ser um produto, um evento ou um indivíduo. Mais especificamente, a análise de sentimentos está preocupada em analisar se um determinado texto está falando algo positivo ou negativo de uma determinada entidade. Neste contexto surgiram estudos com o objetivo de procurar soluções e construir sistemas para facilitar esse processo, o qual é responsável por varrer inúmeros *websites*, buscar opiniões e apresentar os resultados de forma clara para facilitar a tomada de decisão por parte do usuário.

Embora ainda seja uma área em expansão, alguns sistemas automáticos já foram desenvolvidos, como o *Sentiment140*® e o *Opsys*®. Entretanto, a maioria desses trabalhos não é de uso comercial, ficando restritos na maioria das vezes, apenas aos laboratórios de pesquisas.

Um processo de análise de sentimento basicamente é composto pelas seguintes etapas: a escolha da fonte e a busca dos dados relativos à entidade de interesse na análise, o processamento destes dados, a análise dos textos em si e a apresentação dos resultados. Também deve ser observado, segundo SILVA, LIMA, BARROS, (2012), o

fato de que a análise de sentimentos pode ser realizada em três níveis. O primeiro nível é a análise de documentos, no qual o sentimento global expresso no texto é levado em consideração; o segundo nível é a análise de sentenças, no qual é classificada cada sentença do texto; e, por último, tem-se a análise de características, na qual as opiniões são expressas com base nas características dos objetos, como o ambiente de um restaurante, o sabor da comida, o serviço de garçom, etc.

2.1 – As Fontes de Dados

A escolha da fonte de dados para análise de sentimentos sempre vai depender do tipo de aplicação que se pretende desenvolver. Se a intenção é coletar comentários do público em geral que tenha acesso a *Web*, um dos locais que possui uma grande quantidade de informações são as redes de relacionamento social, como o *Facebook*®, o *Google+*® e o *Twitter*®. Geralmente estas redes sociais possuem *APIs* para possibilitar o acesso aos seus dados, sendo esses dados, normalmente, os comentários produzidos pelos usuários. Basicamente, o sistema consumidor passa alguns parâmetros previamente especificados e tem como retorno os comentários, geralmente no formato *JSON* (www.json.org).

No entanto, muitos *websites* e *blogs* são menos estruturados que as páginas de uma rede social, não possuindo um padrão e estruturas compreensíveis para os dados, e nem mesmo *APIs* para sua consulta, tornando árdua esta tarefa. Uma das formas de resolver o problema da falta de *APIs* é por meio da utilização de *webscrapers*, que são softwares especificamente construídos para varrer uma página *Web* extraindo informações úteis para o sistema em questão. O problema com os *webscrapers* é que eles dependem da estrutura física da página *Web*, isto é, de como os elementos *HTML* estão estruturados na página. Infelizmente, essa estrutura varia muito em função da organização da informação que se deseja para o *website*, o que a torna muito instável.

Além dos *webscrapers*, quando se deseja obter informações específicas na *Web* em *websites* que não são bem estruturados como as redes sociais, ou *websites* comerciais consagrados, é necessário utilizar-se de *webcrawlers*. Estes são softwares responsáveis por varrer a *Web* automaticamente em busca de páginas que contém os dados requeridos e de salvá-las para posteriormente processamento. Os sinônimos mais

comuns para este tipo de serviço são *webspider* ou *webrobot*. Alguns exemplos de *webcrawlers* disponíveis são: *Methabot* (http://sourceforge.net/projects/methabot/), *Pavuk* (http://www.pavuk.org/) e *WebSPHINX* (http://www.cs.cmu.edu/~rcm/websphinx/).

2.2 – Dificuldades Encontradas e Pontos a Ser Considerados para a Análise de Sentimentos

A análise de sentimentos tem diversas dificuldades envolvidas nas fases do seu processo. Abaixo são listadas algumas mais frequentes, porém não exclusivas deste trabalho:

- O uso de dialetos regionais. Como a análise é feita com base em textos livres, i.e. escrito por qualquer usuário, não é seguido um padrão linguístico único, causando dificuldades na extração dos termos a serem analisados. Isso ocorre, muitas vezes, pelo uso de dialetos que se altera de acordo com a cultura e local no qual os usuários estão situados;
- Ambiguidade dos comentários ou termos. Esse problema é muito comum, pois inúmeros termos possuem diversos significados, causando equívoco no resultado apresentado ao usuário;
- Detecção de ironias nos comentários. O uso de ironia torna difícil a identificação da qualidade do comentário, pois normalmente o texto inverte o sentido dos termos usados.
- O tamanho do texto a ser analisado. Quando existe limitação no tamanho do texto, como no caso do *Twitter*® que permite somente 140 caracteres, os usuários criam e utiliza-se de muitas abreviaturas e termos próprios para expressar suas opiniões, o que geralmente não é classificado como termo opinativo pelo classificador.

Além dos pontos enumerados acima, a análise de sentimentos possui uma série de fatores menores que devem ser levados em consideração para não gerar resultados falsos ao usuário. De acordo com GUTEMBERG, (2010) outro problema muito comum é indicar se um texto é de caráter opinativo ou se somente é um fato, o qual ele considera uma tarefa mais difícil do que a realização da própria análise.

2.3 – Aplicações com Análise de Sentimentos

As aplicações que realizam a análise de sentimentos não estão restritas apenas à avaliações de comentários de produtos e/ou serviços, elas podem também ser utilizadas para outros fins. Abaixo alguns exemplos de aplicações:

- Análise em tempo real: O *Twitter sentiment analysis (http://smm.streamcrab.com/)* é uma aplicação que fica buscando novos comentários em tempo real no *Twitter*®, a partir de um termo informado pelo usuário, e fica atualizando os resultados dos sentimentos. Ele realiza a busca de qualquer termo digitado;
- Observatório da Web: Desenvolvido por pesquisadores da UFMG e parceiros, este projeto consiste em uma ferramenta gratuita dedicada ao monitoramento de importantes fatos, eventos e entidades na rede mundial de computadores em tempo real (http://observatorio.inweb.org.br/). Pesquisa Política: O *Eleitorando* (www.eleitorando.com.br/) é uma aplicação desenvolvida exclusivamente para buscar e analisar comentários de candidatos nas redes sociais em tempo real.

As aplicações citadas acima são produtos que atualmente estão em uso. Mas pensado em projetos que podem ser desenvolvidos e em algo mais genérico podem-se também citar algumas:

- *Reviews* em *websites*: Um sistema desta categoria pode funcionar como um motor de busca, pesquisando em comentários de *websites* opiniões positivas e negativas sobre produtos (PANG e LEE, 2008);
- Aplicação para governança de empresas: O principal foco de qualquer usuário que utiliza um sistema de análise de sentimentos é a tomada de decisão. Pang e Lee (2008) em sua obra *"Opinion Mining and Sentiment Analysis"* traz a seguinte situação: porque os consumidores compram um *notebook* da empresa X e não da empresa Y, sendo que os produtos possuem características semelhantes. Para responder tal questão é necessário desenvolver um sistema que verifica se o produto da empresa Y possui muitos comentários negativos, e procurar pelos comentários do produto

equivalente da empresa X e então, a partir deste ponto, verificar se estão falando algo positivo do produto;

- Ordenação de produtos ou resultados de pesquisas: Geralmente os produtos em *websites* de compra são ordenados pelo menor preço, maior preço e assim por diante. O que pode ser feito é ordenar uma listagem de produtos de acordo com os comentários positivos ou negativos, pois muitas pessoas não estão preocupadas em comprar produtos pelo preço, mas sim pela qualidade e aprovação de outras pessoas. Ou mesmo um motor de busca pode ordenar os resultados de acordo com o quanto positivo são os textos ou comentários de usuários sobre a página.

2.4 – Etapas da Análise de Sentimentos

O desenvolvimento completo de um sistema de classificação de sentimentos é considerado complexo, por isto ele geralmente é dividido em etapas, e muitos trabalhos concentram-se apenas no aperfeiçoamento de uma dessas etapas, como o trabalho apresentado por LIMA, (2011), o qual tem como objetivo central a extração de características de entidades (i.e. se consideramos a entidade "pizza", uma das características positivas seria ela estar quente e uma negativa estar gelada).

Basicamente, um sistema considerado completo, é composto pela coleta, processamento, análise e apresentação de resultado MATIOLI, (2010), como mostra a Figura 1.

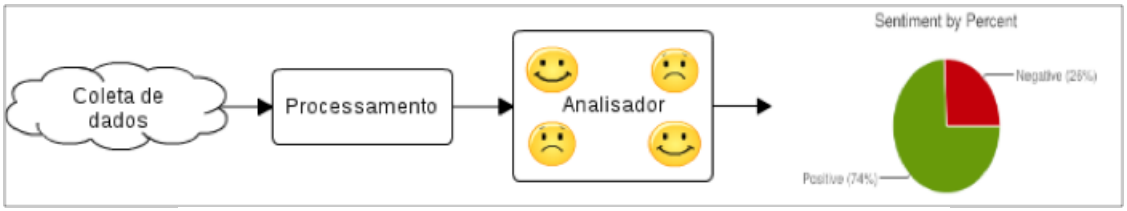

Figura 1 – Etapas comum de um analisador de sentimentos.

Estas etapas se caracterizam por:

- **Coleta dos dados**: nesta etapa é definida qual será a fonte de dados utilizada, a qual pode ser textos da *Web*, comentários, *blogs*, textos de *RSS*, entre outras;

- **Preparação dos dados**: esta etapa é conhecida também como pré-processamento, sendo a etapa na qual os dados passam por um tratamento para tentar corrigir os problemas com a escrita, abreviaturas, gírias, ditos popular, etc. Nesta etapa muitos sistemas também realizam outras operações, como no caso deste trabalho no qual é realizada a tradução dos comentários;
- **Classificação dos sentimentos**: é a etapa principal de um sistema deste tipo. É nela que as técnicas de análise são aplicadas e os textos são classificados como positivo, negativo ou neutro. Esta classificação pode ser realizada utilizando-se de algoritmos de aprendizagem de máquina ou técnicas baseada em recursos léxicos.
- **Sumarização dos resultados**: esta é a última etapa de um sistema de análise de sentimentos. E importante que os resultados sejam mostrados de forma clara para que o usuário possa compreender facilmente. Geralmente os resultados são exibidos em forma textual ou em gráficos. A apresentação dos resultados em forma textual tende a deixar os usuários confusos, devido à abundância de dados, por este motivo a apresentação em gráficos é mais utilizada, principalmente quando os dados são de caráter estatístico, devido à fácil compreensão.

A Figura 2 mostra o resultado apresentado por um sistema deste tipo, a palavra pesquisada para o exemplo foi "obama", utilizando o sistema *Sentiment140*®.

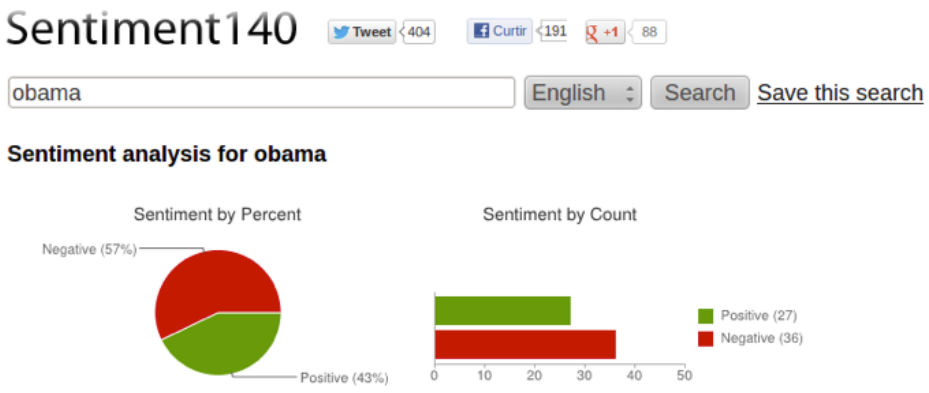

Figura 2 - Resultado apresentado pelo Sentiment140®

Outra maneira de apresentação dos resultados que também é utilizada é mostrada por (PANG e LEE, 2008). Ela possibilita a sumarização de resultados de vários

produtos em um mesmo gráfico, facilitando a visualização e comparação por parte dos usuários, como mostra a Figura 3;

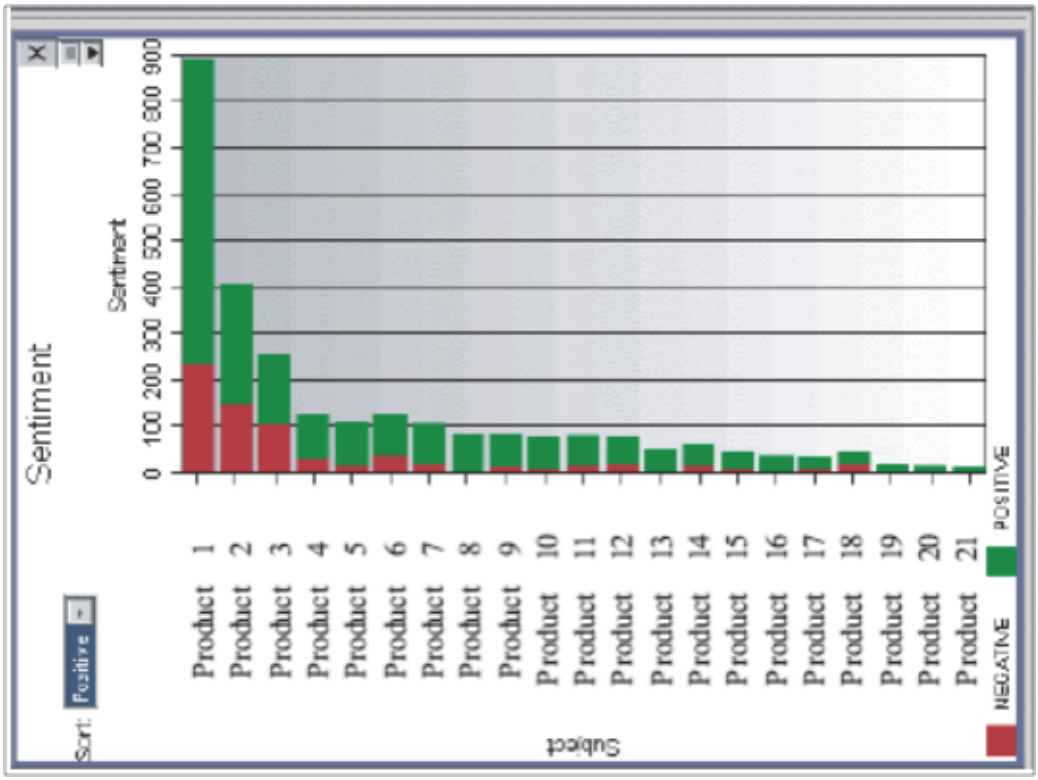

Figura 3 - Apresentação de resultado proposto por (PANG e LEE, 2008).

Neste último caso, o sistema vai armazenando um histórico de pesquisa do usuário e no momento que for solicitado a visualização o conjunto de resultados é exibido.

2.5 – Técnicas Baseadas em Recursos Léxicos

Esta técnica geralmente utiliza uma base que contém termos previamente classificados, a qual associa a cada termo notas positivas ou negativas. Neste contexto cada termo presente no texto a ser analisado é comparado com os termos da base léxica e o seu retorno somado ao valor final do texto. Quando este valor for maior que zero, o sentimento será positivo e se for menor, será negativo. Para uma melhor compreensão, suponha que uma base X é composta pelas palavras “excelente”, “bom” e “certeiro”,

com notas positivas de valor +1 e "ruim" com nota negativa de valor -1. Observe o exemplo da seguinte frase e a análise realizada em nível de documento:

O Messi é um excelente (+1) jogador, ele tem um chute certeiro (+1).

Neste caso temos duas palavras encontradas na base X que possuem valores positivos, por isto a frase é considerada positiva, com valor +2. É interessante observar que no momento da análise o contexto do texto a ser analisado não é levado em consideração, apenas as palavras em si.

Agora observe a frase a seguir:

Aquele jogador não tem nada de ruim (-1), e o jogo foi bom (+1).

Considerando a base X, a palavra "ruim" é considerada negativa, porém devido à presença da palavra "não" ela possui um sentido positivo para quem está lendo Em seguida a frase "mas o jogo foi bom" tem sentido positivo. Somando os valores das duas palavras encontradas na base temos a nota final 0 e, consequentemente, o texto será considerado neutro pelo analisador, contrário ao seu real valor que deveria ser positivo, pois a negativa presente na primeira frase inverte o valor da palavra "ruim". A despeito destes tipo de problemas, a grande vantagem deste tipo de abordagem é o fato de que a análise pode ser realizada sem que o domínio de negócio seja levado em consideração.

2.6 – A SentiWordNet©

A *SentiWordNet*© é uma base léxica, em formato de arquivo de texto, utilizada na análise de sentimentos. Ela foi desenvolvida utilizando a base de palavras da *WordNet*© (http://wordnet.princeton.edu/), a qual é composta por substantivos, verbos, adjetivos e advérbios agrupados em um conjunto de sinônimos cognitivos (*synsets*), cada um expressando um conceito distinto (http://wordnet.princeton.edu/). Atualmente, a *SentiWordNet*© está na versão 3.0 e é publicamente disponível para fins de pesquisas.

A *SentiWordNet*© é resultado da anotação automática de cada *synset* da *WordNet*©, na qual foi atribuído a cada termo dois valores numéricos, indicando o *Pos* (Positivo) e *Neg* (Negativo). O valor neutro é implícito na base do *SentiWordNet*©,

sendo necessário aplicar a equação **Obj = 1 – (Pos+Neg)** para obter o valor correspondente. O valor máximo somando as três classificações sempre será igual a 1. Outro ponto a ser considerado é que dependendo do contexto que um termo é utilizado o seu valor é diferente. A Figura 4 mostram um exemplo.

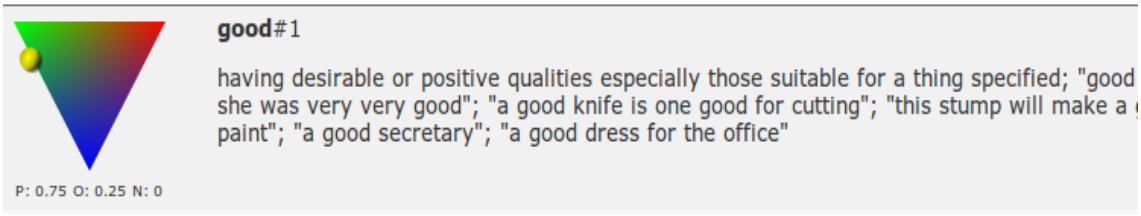

Figura 4 - Resultado de pesquisa no website SentiWordNet©.

Nos dois caso a palavra *good* aparece como adjetivo. Na Figura 4 ela tem como valor *Pos*: 1, *Neg*: 0 e *Obj*: 0 e na Figura 5 o *Pos*: 0,625, *Neg*: 0 e *Obj*: 0,375.

A *SentiWordNet*© mantém uma interface *Web* para disponibilizar consultas aos termos, retornando todos os resultados disponíveis. Já para utilização no sistema analisador é necessário fazer o *download* de um arquivo texto disponível. A Tabela 1 mostra um exemplo deste arquivo.

Tabela 1 - Estrutura da *SentiWordNet*©.

Classe Gram.	***Id***	***Valor Pos.***	***Valor Neg.***	***Termos/Qtd. Ocorrências***	***Frase***
Adjetivo	00059028	0	0.125	inflamed#3	Adorned with tongues of flame
Substantivo	08178741	0	0	population#2	A group of organisms of the same species inhabiting a given area;"they hired hunters tokeep down the deerp opulation"
Advérbio	00370046	0.25	0	incisively#1	In anincisive manner; "hew as incisively critical"
Verbo	02737724	0.75	0	belong#2	Be suitable or accep table; "This students ome how doesn't belong"

3 – DESCRIÇÃO DO SISTEMA DESENVOLVIDO

Para se desenvolver um projeto de análise de sentimentos de comentários, textos, notícias, entre outros, é necessário estudar os modelos e técnicas disponíveis e verificar quais são os mais úteis ao tipo de problema a ser resolvido. Por exemplo, é importante avaliar se a análise vai ser restrita a um determinado domínio ou não, pois isto ajuda a determinar as técnicas mais apropriadas.

Para a classificação de sentimentos este trabalho baseou-se na abordagem de DENECKE, (2008), a qual utiliza a *SentiWordNet*© por meio da tradução do texto original a ser analisado para a língua inglesa, a língua nativa da *SentiWordNet*©, e, então, faz a aplicação do algoritmo para realizar a classificação das palavras do texto.

Outro ponto a ser considerado é a escolha da fonte da qual os dados serão obtidos. Considerando o aumento de usuários nas redes sociais e o consequente aumento da quantidade de informações compartilhadas, neste trabalho foi escolhida a utilização da rede social *Twitter*® para servir como fonte de dados.

Tendo em vista que não foi encontrado trabalho que aborde a utilização da *SentiWordNet*© especificamente para a língua portuguesa, este trabalho tem por objetivo utilizar a *SentiWordNet*©, juntamente com um mecanismo de tradução da língua português para a inglesa, para classificar comentários em português.

Um fator que prejudica muito os resultados apresentados por um sistema deste tipo, i.e. que analisa principalmente redes sociais, é o uso de termos coloquiais por parte dos usuários. Com intuito de tentar resolver esse problema foi criado um dicionário de gírias e abreviaturas, como detalhado a seguir.

Para maior disponibilidade e facilidade de acesso, o protótipo possui uma interface *Web*, tendo sido desenvolvido utilizando-se da linguagem de programação *Java*.

3.1 – Arquitetura Geral do Sistema

A Figura 5 ilustra a arquitetura geral do protótipo desenvolvido. Os módulos que compõe o sistema são: a busca, o pré-processamento, o analisador e a sumarização de resultados.

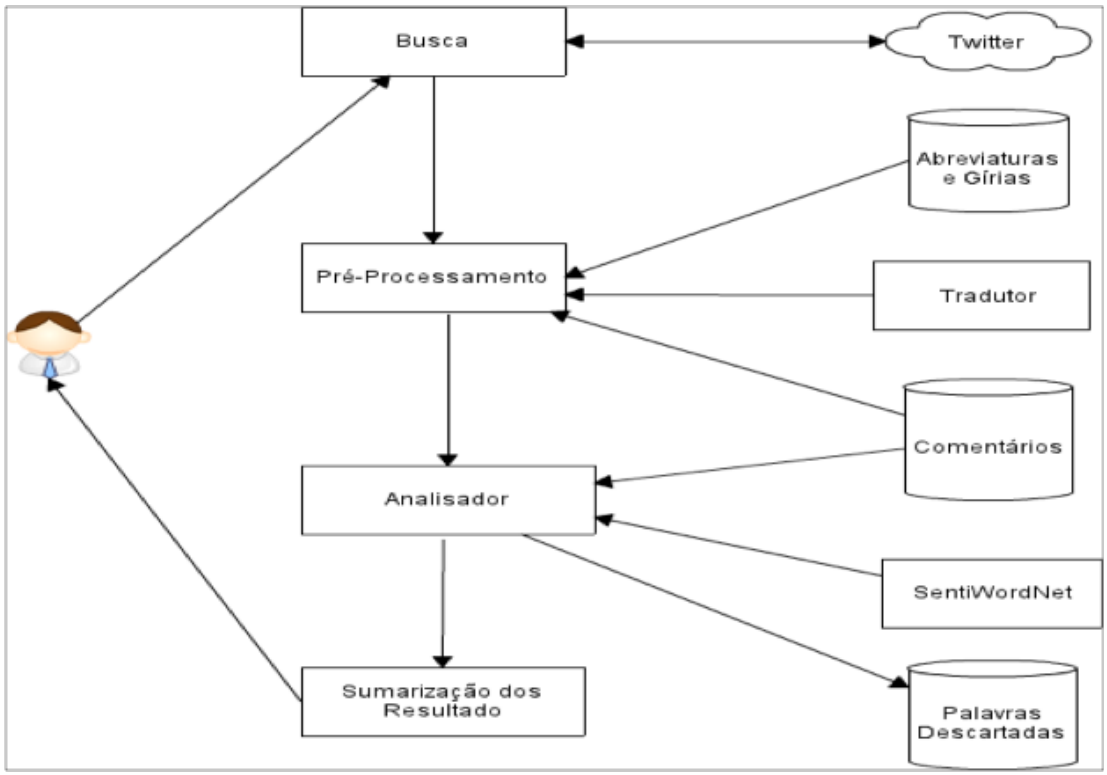

Figura 5 - Arquitetura do projeto desenvolvido.

Para o funcionamento do sistema é necessário que o usuário forneça o nome da **entidade** que ele quer pesquisar. A partir deste nome o módulo de busca é acionado fazendo uma solicitação ao sistema *Twitter*®, o qual, em geral, retorna em torno de noventa a cento e dez *tweets* para cada pesquisa.

Para a avaliação das opiniões contidas nos *tweets* é necessário fazer um pré-processamento, no qual são aplicadas técnicas para deixar o texto mais interpretável para o analisador. Devido à grande utilização de abreviaturas e gírias por parte dos usuários, o sistema consulta um dicionário de gírias e abreviaturas, mantido pelo próprio sistema, fazendo a injeção de palavras completas para tentar minimizar este problema.

O módulo analisador é o centro do sistema, sendo responsável por classificar as opiniões expressas no *tweets* em português como positivas, negativas ou neutras. Para realizar esta classificação foi utilizada a *SentiWordNet*© a qual contém milhares de *synsets* com dois valores de opiniões atribuído a cada um. Como este arquivo está na língua inglesa foi necessária a utilização de um tradutor do português para o inglês. As palavras que não forem encontradas na *SentiWordNet*© serão armazenadas para análise futura.

Após análise dos comentários, a próxima etapa do sistema é responsável pela apresentação dos resultados ao usuário que, de acordo com a literatura estudada, pode ser realizado de forma textual ou na forma gráfica. Neste trabalho, por se tratar de um protótipo, foram utilizados recursos textuais para a sumarização dos resultados, porém sistemas comerciais como o *Sentiment140*® utilizam-se mais de gráficos, por tornar mais fácil a compreensão por parte do usuário.

3.2 – O Módulo de Busca

Este módulo tem a responsabilidade de consultar a fonte dos dados, no caso o *Twitter*®, e extrair os comentários para que seja realizada a análise de sentimentos. Os comentários são extraídos de forma automática, por meio do nome de uma entidade informada pelo usuário. Essa entidade é o objeto sobre o qual se deseja saber as opiniões, podendo ser um produto, pessoa ou até mesmo um serviço.

Basicamente a extração de comentários da rede de informações *Twitter*® pode ser realizada de duas maneiras. A primeira forma é a *Twitter Search API* (https://dev.twitter.com/docs/api/1/get/search), uma *API* disponível no próprio *website* do serviço, a qual disponibiliza os dados no formato *JSON* (www.json.org/), o que facilita seu processamento por diversas linguagens, sendo necessária somente a leitura do arquivo no formato texto.

Para realizar a consulta é preciso montar uma *query* a partir dos parâmetros disponíveis. A Tabela 2 mostra alguns exemplos.

Tabela 2 - Alguns parâmetros aceito pela API do Twitter®.

Parâmetros	***Descrição***	***Requerido***
q	Parâmetro para realizar a consulta	sim
result_type	Tipo de resultados da pesquisa “mixed”: retorna comentários recentes e populares “recent”: retorna comentários postados recente “popular”: retorna comentários populares	não
locale	Especifica qual idioma será pesquisado	não

O código abaixo mostra um exemplo da *URL* de requisição e o resultado.

```
GET http://search.twitter.com/search.json?q=Twitter20API&result_type=mixed
{ "created_at": "Wed, 19 Jan 2011 19:55:10 +0000", profile_image_url":
"http://a0.twimg.com/profile_images/585494683/13038_613894593395_
24403188_35452430_7524658_n_normal.jpg",
    "from_user_id_str": "1493373",
    "id_str": "27816354073550848",
    "from_user": "mirandafte",
    "text": "@mirandafte testing some twitter API - ness",
    "to_user_id": 1493373,
    "metadata": {
      "result_type": "recent"
    },
    "id": 27816354073550848,
    "geo": null,
    "to_user": "mirandafte",
    "from_user_id": 1493373,
    "iso language code": "en",
    "source": "&lt;a href=&quot;http://twitter.com/&quot;&gt;web&lt;/a&gt;",
    "to_user_id_str": "1493373"
  }
```

A segunda forma faz uso da linguagem *Java*, para qual é disponibilizada uma biblioteca *open source* que realiza o acesso a *API* do *Twitter®*. Segundo WINTERWELL, (2011) essa biblioteca é classificada como robusta e de fácil uso. No caso da consulta, uma das vantagens de sua utilização é o fato de o seu retorno ser um objeto, característica comum dentro da linguagem *Java*.

O código abaixo mostra um exemplo de código da utilização da mesma.

```
Twitter t = new Twitter();
      List<Status> listas = t.search("maringa");
      for (Status s : listas) {
         System.out.println("usuario: " + s.getUser());
```

```
System.out.println("Mensagem: " + s.getText());  }
```

Para o presente trabalho escolheu-se esta última abordagem, devido ao fato do desenvolvimento da aplicação se dar em cima da linguagem *Java*.

A partir do momento que o sistema obtém o resultado da pesquisa, todos os comentários são armazenados em um banco de dados em seu formato original, bem como o termo que o usuário digitou e a data em que a pesquisa foi realizada. Todos os comentários são armazenados no banco para que futuramente possam ser analisados possíveis erros com a tradução e o processamento do texto, buscando outras formas de melhorar a classificação.

3.3 – O Módulo de Pré-Processamento

A etapa de extração de texto é comum a quase todos os sistemas que trabalham com processamento e análise de textos, principalmente quando os mesmo são retirados da *Web*, na qual a utilização de linguagens informais é muito utilizada.

Esta é uma das etapas mais importantes de um sistema de análise de sentimentos, pois é quando se realiza toda a preparação dos dados para futura análise.

3.3.1 – Dicionário de Abreviatura e Gírias

Como já mencionado anteriormente, um dos maiores problemas enfrentados atualmente na área de análise de sentimentos é a forma com que os usuários escrevem, utilizando-se de abreviaturas, gírias e linguagens informais. Muitas vezes o usuário sente-se forçado a usar tais linguagens, pois, por exemplo, o *Twitter*® permite somente *posts* com no máximo 140 caracteres. Na tentativa de resolver tal problema, este módulo conta com um dicionário de abreviaturas e gírias, criado manualmente pela observação de diversas redes sociais e extraindo os termos informais utilizados e que pôde ser observado. A base foi construída de forma ad hoc, e conforme os termos foram aparecendo.

A primeira etapa do pré-processamento é a troca das abreviaturas e gírias pela forma extensa da palavra. O exemplo da Tabela 3 mostra a estrutura e alguns termos do

dicionário, sendo que na primeira coluna encontra-se a palavra na forma original, na segunda seguida pelo delimitador # a abreviatura/gíria. A final tem-se uma frase retirada do *Twitter*® na forma original e após a substituição dos termos.

Tabela 3 - Estrutura do dicionário de abreviaturas e siglas.

não	#ñ
mensagem	#msg
tambem	#tmb
você	#vc
brasil	#br
Exemplo aplicação base: **Antes:** vc gostou do br jogando? **Depois:** você gostou do brasil jogando?	

3.3.2 – Tradução

Para realizar a análise de sentimentos de comentários na língua portuguesa inicialmente é necessário que os mesmos sejam traduzidos para o inglês, que é a língua oficial da *SentiWordNet*©. No entanto, sabe-se que muitos textos podem até mesmo perder o sentido devido à falta de contexto no momento da tradução.

A tradução foi realizada pelo serviço de tradução *Google Translate*® (translate.google.com.br/), um dos mais conhecidos atualmente, fornecendo suporte a tradução para inúmeras línguas. A tradução de textos do *Google Translate*® recusa alguns documentos devido a alguns caracteres especiais, um exemplo é o #, o qual deve ser retirado do texto. Outro problema nele encontrado é a substituição dos espaços em branco pela sequência %20.

Como o serviço de *API* não é gratuito, sendo cobrado pela quantidade de caracteres traduzidos, algumas partes dos comentários que não influenciam no resultado da análise foram removidas no momento do processamento de textos como, por exemplo, *links* para páginas na *Web*.

3.3.3 – O Stemming

O processo de *stemming* tem como objetivo diminuir as palavras a sua raiz, reduzindo, assim, o número de termos a serem processados. Por exemplo, se em um texto se encontrar o termo "*cars*" ele é convertido para "*car*". Essa etapa é necessária porque caso a palavra no plural acima citada for procurada na *SentiWordNet*© nada será retornado, pois a *SentiWordNet*© não contém plurais.

Para realizar tal tarefa foi utilizado neste trabalho o *Apache Lucene* (http://lucene.apache.org/core/), o qual é um *framework* usado principalmente para a indexação de documentos, mas que contém várias funções para o processamento/extração de textos de documentos *Web*. Dentre estas funções existe uma que realiza o processo de *stemming* para a língua inglesa, e esta parte é utilizada neste trabalho. O *stemming* é realizado somente no momento da busca da palavra no *SentiWordNet*©. Esta etapa não é aplicada na frase como um todo, de forma que ela é armazenada no formato original após a tradução, possibilitando, assim, processamentos futuros. Após o processamento de cada etapa que compõe este módulo a frase é atualizada no banco de dados.

3.4 - A Classificação dos Comentários

Para aferir a taxa de acerto do sistema, é necessária antes da classificação automática a realização de uma classificação manual dos comentários para servir de base de comparação.

Um motivo para que a classificação manual seja sempre anterior a do sistema é para que a resposta do sistema não venha influenciar quem a estiver realizando. Outro motivo, é que o avaliador, tem que conhecer o domínio de negócio do tema pesquisado. No caso desta base, a classificação foi realizada pelo próprio autor. Por exemplo, se uma pessoa que não tem o mínimo conhecimento de futebol for avaliar o comentário da Figura 6, ela pode não saber o que é *série B*.

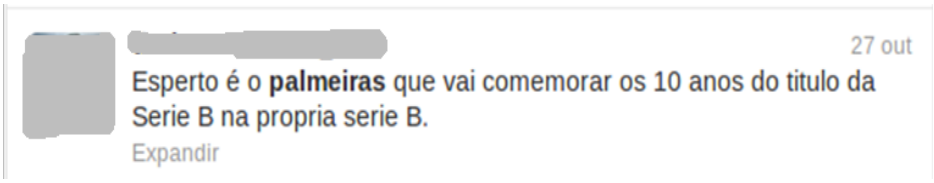

Figura 6 - Comentário no Twitter®.

Adicionar Mensagem | Totalizar | **Total Termos: 3 Acertos: 3 Positivos Humano: 0 Negativos Humano: 0 Neut Neutro Sistema: 0 Acertos Pos.: 0 Acertos Nega.: 0 Acertos Netro: 0**

Lista das mensagens

ID	Frase Original	Frase Final	Frase Ingles
215	Não trate como viagem à Disney quem te trata como excursão ao Beto Carreiro.	não trate como viagem à disney quem te trata como excursão ao beto carreiro.	no treat for trip to disney who treats you like the tour beto path.
216	Preparem-se para o JUMP2013 no Beto Carreiro World! http://t.co/mVcygTgf Aguardem, em breve mais informações! Será PODER! Quem vai? RT	preparem-se para o jump2013 no beto carreiro world! em breve mais informações! será poder! quem vai? rt	prepare for the jump2013 the path beto world! more information coming soon! will power! who will? rt
217	Esperando acabar o jogo do timão arrumar minhas coisas, pq amanhã Beto Carreiro	esperando acabar o jogo do timão arrumar minhas coisas, pq amanhã beto carreiro	hoping to finish the game the wheel pack my things, 'cause tomorrow beto pa

Figura 7 - Resultado pós Processamento.

A Figura 7 mostra uma tela do sistema após as etapas de processamento de texto, tradução e da análise manual realizada, com os textos prontos para realizar a análise.

3.5 – O Módulo Analisador

Este módulo é considerado o centro do sistema, pois é nele que a classificação automática dos comentários é realizada. Como dito anteriormente, esta tarefa foi realizada utilizando-se da linguagem de programação *Java* e da base de palavra da *SentiWordNet*©, a qual retorna diversos significados para uma mesma palavra. Por exemplo, a palavra *good* possui como adjetivo 27 significados diferentes, 4 como substantivos e 2 como advérbios, totalizando 33 retornos.

Para calcular a nota de cada palavra contida no texto é utilizada a Equação 1, na qual a variável ***si*** é o *score* de cada significado retornado do *SentiWordNet*© e ***t*** é o tamanho do conjunto dos *scores* individuais SOUZA, (2011), por exemplo, o termo *good*, pode possuir 2 ocorrências, ora como adjetivo e ora como advérbio.

$$Score = \sum_{i=0}^{t} \frac{1}{i+1} * si \quad Soma = \sum_{i=1}^{t} \frac{1}{i} \quad Score\ final = \frac{Score}{Soma}$$

Equação 1- Cálculos dos scores individuais.

Ao final se obterá o *score* final da palavra, no qual:

1. Se o *score* final > = 0,75, a palavra é considerada fortemente positiva.
2. Se o *score* final > 0,25 e < 0,75, é considerado positivo.
3. Se o *score* > 0 e < = 0,25, a palavra é fracamente positiva.
4. Se o *score* = 0 a palavra é neutra.
5. Se o *score* < 0 e > = -0,25 é considerado pouco negativo.
6. Se o *score* < -0,25 e > = -0,75 é considerado negativo.
7. Se o *score* for < -0,75 é considerado muito negativo.

Estes dados servem para verificar o quanto cada palavra pode influenciar nos textos.

Com cada palavra que compõe o texto classificada, tem-se que verificar o sentimento de todo o documento, sendo necessário somar todos os *scores* finais individuais. Por fim é feita uma análise, sendo considerado *negativo* os textos que assumirem um valor menor do que 0 e, consequentemente, *positivos* os aqueles que assumirem valor maior que zero, e neutro os que o *score* final é igual a zero.

Assim como o pré-processamento, o resultado da análise e o valor da avaliação do comentário são armazenados na base de dados, permitindo consultas posteriores. A Figura 8 mostra a imagem de uma tela do sistema.

Adicionar Mensagem | Totalizar **Total Termos: 100 Acertos: 46 Positivos Humano: 14 Negativos Humano: 76 Neutro Humano: 10**
Sistema: 45 Neutro Sistema: 6 Acertos Pos.: 8 Acertos Nega.: 38 Acertos Netro: 0

Lista das mensagens

ID	Frase Original	Frase Final	Frase Ingles	Sentimento Hum.	Sent. Sist.
7	tim. vc sem fronteiras, sem sinal, sem conseguir falar com ninguem, sem internet 3g no celular etc	tim. vc sem fronteiras, sem sinal, sem conseguir falar com ninguem, sem internet 3g no celular etc	tim. vc no boundaries, no signal, unable to talk to anyone, without 3g internet on mobile etc.	Negativo	NEGATIVO
8	Existe algo pior que o 3G da Tim! Eu não sei o que, mas deve existir...	existe algo pior que o 3g da tim! eu não sei o que, mas deve existir.	There is something worse than 3g tim! I do not know what, but it must exist.	Negativo	NEGATIVO
9	Brasil fica no top 10 dos países que mais utilizam smartphones com Android ou IOS e rede 3G!	brasil fica no top 10 dos países que mais utilizam smartphones com android ou ios e rede 3g!	Brazil is in the top 10 countries that use more smartphones with android or ios and 3g network!	Positivo	POSITIVO

Figura 8 - Comentários com as classificações.

3.6 – Tratamento das Palavras não Encontradas

As palavras que não forem encontradas na *SentiwordNet*© são armazenadas em um arquivo para que manualmente seja feita uma análise e verificado o motivo do descarte.

Verificando esta base de palavras, pôde-se observar que a maior causa de não retorno de valores foi devido à linguagem informal utilizada pelos usuários, outra causa foram palavras comuns na língua portuguesa que não foram traduzidas como, por exemplo, a palavra *otimo* (sem acento).

O que pode ser feito com muitos descartes é substituí-los no momento anterior a tradução, adicionando a palavra no dicionário de abreviaturas e siglas. Utilizando o exemplo da palavra *ótimo*, a mesma pode ser substituída por *bom* ou acentuá-la, desde que permaneça o mesmo sentido da frase sem prejuízo do resultado.

4 –Avaliação do Protótipo Proposto

Neste capítulo são descritos a metodologia empregada na realização dos teses realizados para a avaliação do protótipo proposto, assim como os resultados obtidos e os problemas encontrados.

4.1 – Metodologia de Avaliação

Oficialmente, não foi encontrada nenhuma base de comentários para a língua portuguesa com suas respectivas classificações para realizar os testes de desempenho do protótipo desenvolvido. Portanto foi criada uma base específica para este trabalho, a qual foi classificada manualmente. Esta base contém comentários extraídos principalmente do *Twitter*®, alguns de vídeos do *Youtube*® e classificações de vendedores do *Mercado Livre*®.

A base conta com aproximadamente 1000 comentários. Para tanto, foram realizadas várias consultas no *Twitter*®, com retornos de 100 comentários cada. As consultas foram realizadas utilizando diferentes termos, sempre procurando diversificar os domínios de negócios como, por exemplo, marcas de produtos, locais, pessoas, etc. Isto é importante, pois pôde ser observado que dependendo do domínio de negócio o público que faz os comentários, além de ser diferente, utiliza também uma linguagem característica do domínio para se expressar, o que sempre altera o resultado. Todas as consultas do *Twitter*® foram realizadas utilizando a interface do sistema.

Em uma segunda etapa foi feita a inclusão de comentários do *Youtube*® e do *Mercado Livre*® para comparar os diferentes tipos de ambientes. Isto foi realizado pois no *Twitter*® muitos comentários não possuem caráter opinativo, e a maioria das vezes não diz respeito ao termo pesquisado, diferente do que ocorre no *Youtube*® e *Mercado Livre*®, nos quais o usuário escreve quase que sempre um comentário opinativo sobre o objeto principal. Os comentário do *Youtube*® e Mercado Livre foram incluídos

manualmente A tabela 4 mostra os termos que foram usados nas pesquisas para construção da base.

Tabela 4 - Termos pesquisados.

Termo Pesquisado	***Área***	***Fonte***
Beto Carreiro	Diversão	*Twitter*®
3G	Tecnologia	*Twitter*®
Boticário	Beleza	*Twitter*®
Fernando Haddad	Política	*Twitter*®
Jose Serra	Política	*Twitter*®
Palmeiras	Clube de Futebol	*Twitter*®
Flamengo	Clube de Futebol	*Twitter*®
Busca Implacável 2	Filme	*Twitter*®
Neymar	Pessoa	*Twitter*®
Golf	Carro	*Twitter*®
Terror Sobre Rodas – Negativos	Filme	*YouTube*®
Falcão Negro em Perigo – Positivos	Filme	*YouTube*®
Roupa Nova 30 Anos – Positivos	Grupo Músical	*YouTube*®
Comentários vendedor – Positivos	Pessoa	*Mercado Livre*®
Comentários vendedor Negativos	Pessoa	*Mercado Livre*®

A inclusão de comentários do *Youtube*® e do *Mercado Livre*® também está relacionada ao fato de que a maioria das bases de testes encontradas foram construídas utilizando *websites* de *reviews*, porém todos para a língua inglesa. Como exemplo dessas bases pode-se citar o *Cornell movie-review*, disponível no *website* http://www.cs.cornell.edu/people/pabo/movie-review-data, o qual consiste de um conjunto de comentários de filmes composto por 1.000 comentários positivos e 1.000 negativos (PANG e LEE, 2008). Outro Exemplo é o *Multiple-aspect restaurant reviews,* o qual conta com 4.448 comentários, nos quais para cada revisão é dada uma classificação de 1 a 5, em cinco aspectos diferentes: alimentação, ambiente, serviço, valor e experiência geral.

4.2 – Os Testes do Protótipo

Os testes do sistema proposto foram realizados separadamente, ou seja, primeiramente forma realizados testes utilizando os comentários do *Twitter*®, seguido pelos do *YouTube*® e do *Mercado Livre*®.

A primeira fase dos testes foi realizada com os comentários retirados do *Twitter*®, que totalizaram 1000 comentários, os quais, por meio da avaliação manual, foram considerados 54% positivos, 32% negativos e 14% neutros. Esse grupo foi dividido em 10 subgrupos com 100 comentários cada, com um número variável de positivos, negativos e neutros.

A Tabela 5 traz todo o conjunto da classificação manual e a realizada pelo sistema e seus respectivos acertos.

Tabela 5 - Classificações dos comentários retirados do Twitter®.

FLAMENGO						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
57	17	26	51	43	6	40
PALMEIRAS						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
11	84	5	47	51	2	56
GOLF						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
29	4	67	43	34	23	34
BOTICARIO						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
81	10	9	70	24	4	69
BUSCA IMPLACAVEL 2						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
92	7	1	23	75	2	27
JOSE SERRA						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
29	67	6	67	24	9	35
FERNANDO HADDAD						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
79	15	6	52	35	15	54
NEYMAR						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
73	26	1	57	36	7	59

3G						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
14	76	9	51	44	5	43
BETO CARREIRO						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
74	16	9	46	40	14	47

Analisando a possibilidade da utilização do sistema na tomada de decisão, pode-se observar que em apenas três casos houve contradição para mais, ou para menos, de positivos e negativos entre a classificação manual e a automática, sendo eles: *Busca Implacável 2*, *Jose Serra* e *3G*. Isto é um sinal positivo de que poderíamos utilizar este sistema para auxiliar seus usuários a tomar decisões sobre o tema de seu interesse.

A Tabela 6 mostra em percentual o total das classificações realizadas pelo avaliador humano e pelo sistema.

Tabela 6 – Matriz de Confusão – Comentários Twitter®.

Sistema	**Real**		
	Positivo	Negativo	Neutro
Positivo	**299**	145	53
Negativo	203	**146**	67
Neutro	37	31	**19**
Total	**539**	**322**	**139**

Os dados da Tabela 6 mostram os totais da classificação real e os acertos do sistema em cada classe. A partir da tabela foram calculadas as taxas de verdadeiro positivos, verdadeiros negativos e verdadeiros neutro, que representam 55%, 45% e 14%, respectivamente e também os falsos positivos, falsos negativos e falsos neutros, que obtiveram respectivamente 44%, 55% e 86%. Estes resultados mostram claramente uma grande dificuldade em determinar quando uma opinião é neutra.

A média geral de acertos deste conjunto foi de 46,4% com desvio padrão de 12,38%, a qual é um valor baixo, porém dentro da faixa de valores encontrados na literatura para este tipo de fonte de informação. Como os resultados apresentado por DENECKE (2008), no qual as taxas de acertos ficaram entre 51% e 69%, inicialmente esperavam-se os mesmos níveis de acerto para este trabalho, porém em apenas quatro casos as porcentagens de acertos estiveram entre esta média. Alguns pontos foram levantados para tentar entender o que contribuiu para uma baixa taxa de acertos. A

seguir são apresentados alguns casos típicos que podem influenciar no momento da análise. Observe as frases da Tabela 7.

Tabela 7 - Exemplo de opiniões retiradas do Twitter®.

Comentário 1: Hoje é sexta-feira, filme XYZ, uhull. Comentário 2: Fui ao cinema, gostei muito do filme XYZ, ele é muito bom... excelente. Bem melhor que o filme X, que é ruim.

Comentários como estes, i.e. que utilizam esse tipo de linguagem, são muito comuns no *Twitter*®. Se considerarmos o termo "uhull", como uma comemoração pelo fato de que supostamente o usuário irá ver o filme XYZ, o comentário pode ser considerado pela avaliação humana como positivo. Porém para o sistema fica difícil de se chegar a esta avaliação, pois o comentário não possui palavras por ele consideradas positivas.

Agora considere o caso em que o usuário esteja pesquisando pelo filme X. O comentário 2 é considerado de caráter opinativo e o resultado de sua classificação do sistema será positiva aos invés de negativa. Isto ocorrerá porque na última sentença, que realmente fala sobre o filme X, tem um termo que o classifica como negativo. Para problemas como este é interessante considerar a análise de sentenças, o que não foi feito neste trabalho.

Reforçando o que já foi dito anteriormente, os comentários do *Twitter*® muitas vezes são apenas fatos sem valor de opinião, os usuários citam um determinado termo e fazem comentários sem fazer menção das qualidades ou defeitos do termo. A Figura 9 mostra um exemplo de comentário retirado do *Twitter*®.

Figura 9 - Comentários retirado do Twitter®

Nestes casos o termo digitado foi a palavra *golf*, referindo-se ao carro. Começando a análise pelo primeiro comentário, podemos observar que na mensagem em si o termo *golf* não é encontrado, porém o resultado veio na lista de respostas da consulta realizada porque o termo é utilizado como parte de nome do usuário. O segundo resultado refere-se ao jogo de *golf*. Somente o terceiro comentário é que está dentro do contexto do termo intencionado pelo usuário. No primeiro teste realizado, na consulta efetuada utilizando o termo citado acima, 67 ocorrências foram classificadas como neutras. A solução neste caso seria realizar a desambiguação do termo em cada comentário.

Outro ponto que influencia nos acertos, no caso do *Twitter®,* são os *Retweets,* que nada mais são que o compartilhamento de uma mensagem de um amigo para os seus seguidores. Nas consultas realizadas pôde-se notar que em um determinado termo pesquisado em uma lista de 100 comentários, chegou-se a ter 15 comentários iguais. A primeira hipótese a se considerar para solucionar o problema seria excluir os repetidos, mas deve-se levar em conta o fato de que se uma pessoa está compartilhando algo é porque sua opinião a respeito do assunto é a mesma. O problema no momento da análise é que, se o sistema errar na classificação de uma mensagem no grupo por falta de desambiguação do termo, ele automaticamente cometerá 15 erros. Este problema requer um estudo mais profundo para verificar o quanto os *Retweets* são importantes.

Após a análise dos resultados, mediante a baixa taxa de acertos, algumas hipóteses foram levantadas. O fator geral que influenciou diretamente na média final da taxa de acertos foi a fonte de comentários utilizada, neste caso o *Twitter®*, seguido por razões que estão ligadas aos tipos de linguagens utilizadas pelos usuários, que na maioria das vezes são informais. Dos termos que obtiveram baixas taxas de acertos, algumas características podem ser elencadas, como a ambiguidade, o fato de que os comentários realizados no *Twitter®* não estão ligados diretamente a um produto e/ou serviços, o fato de que muitas vezes os usuários citam determinada entidade sem fazer menção opinativa sobre a mesma, e também a utilização de palavras consideradas negativas pelo classificador para comentar algo positivo. Já os termos que apresentaram as taxas de acertos dentro da média esperada, no momento em que os comentários foram extraídos estavam em destaque na mídia, este é um fator que influencia as

pessoas a postarem suas opiniões a respeito, e geralmente são mais claros e objetivos. As taxas de acerto também sobem à medida que a linguagem utilizada é mais formal.

Como o trabalho desenvolvido por DENECKE (2008) realizou seus testes empregando de comentários retirados de *websites* de *reviews* de filmes, para avaliar nosso protótipo de forma mais equitativa realizamos mais alguns testes utilizando-se de fontes equivalentes a dele, no caso o *Youtube*® no Brasil. Além disso, avaliamos também comentários vindos de outro *website* de *reviews*, no caso o *Mercado Livre*®, pois este *website* conta com a classificação do comentário feita pela própria pessoa que o postou, o que é interessante pois retira o viés de uma avaliação posterior indireta.

Nesta fase os testes foram realizados utilizando a base construída exclusivamente com opiniões positivas e negativas, previamente classificadas, que foram retirados do *Youtube*® e do *Mercado Livre*®. A primeira base conta com 300 comentários retirados do *Youtube*®, sendo 200 positivos e 100 negativos. A Tabela 8 mostra o resultado das classificações e acertos.

Tabela 8- Classificação de comentários - Youtube®.

Positivo – Youtube® – Roupa Nova 30 Anos						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
100	**0**	**0**	**95**	**5**	**0**	**95**
Positivo – Youtube® – Filme Terror Sobre Rodas						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
0	**100**	**0**	**21**	**74**	**5**	**74**
Positivo – Youtube® – Filme Falcão Negro em Perigo						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
100	**0**	**0**	**88**	**12**	**0**	**88**

A tabela 9 traz em percentual o total das classificações humana e do sistema.

Tabela 9 – Matriz de Confusão - Youtube®.

Sistema	**Real**		
	Positivo	Negativo	Neutro
Positivo	**183**	21	0
Negativo	17	**74**	0
Neutro	0	5	**0**
	200	**100**	**0**

Para a base de comentários do *Youtube*® pôde ser observado que a taxa de acerto subiu consideravelmente, quando comparado a base de comentários do *Twitter*®. Os

dados da Tabela 9 mostram os totais da classificação real e os acertos do sistema em cada classe. A partir da tabela foram calculadas as taxas de verdadeiros positivos e verdadeiros negativos, que representam 92% e 74%, respectivamente e também os falsos positivos e falsos negativos, que obtiveram respectivamente 8,5% e 26%. Neste caso a média de acertos total foi de 85.67% com desvio padrão de 8,73.

O próximo caso de teste conta com 200 comentários retirados dos *reviews* de compradores do *Mercado Livre*®, sendo 100 considerados positivos e 100 negativos, neste caso os comentários já são classificados pelas próprias pessoas que o realizam. A Tabela 10 mostra o resultado das classificações e de acertos.

Tabela 10 - Classificação de comentários - Mercado Livre®.

Positivo – Vendedores Mercado Livre®						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
100	**0**	**0**	**75**	**19**	**6**	**75**
Negativo – Vendedores Mercado Livre®						
POS. HUM.	***NEG. HUM.***	***NEU. HUM.***	***POS. SIST.***	***NEG. SIST***	***NEU. SIST.***	***ACERTOS***
0	**100**	**0**	**21**	**77**	**2**	**77**

A Tabela 11 traz em percentual o total das classificações humana e do sistema.

Tabela 11 – Matriz de Confusão - Mercado Livre®.

Sistema	**Real**		
	Positivo	Negativo	Neutro
Positivo	**75**	21	0
Negativo	19	**77**	0
Neutro	6	2	**0**
	100	**100**	**0**

Assim como na base de comentários do *Youtube*®, pôde-se notar uma alta taxa de acerto, se comparado ao *Twitter*®. Os dados da tabela 9 mostram os totais da classificação real e os acertos do sistema em cada classe. A partir da tabela foram calculadas as taxas de verdadeiro positivos e verdadeiro negativos, que representam 75% e 77%, respectivamente e também os falsos positivos e falsos negativos, que obtiveram respectivamente 25% e 24%. Neste caso a média geral de acertos foi de 76% com um desvio padrão de 1.

O primeiro questionamento ao se analisar os três testes é a razão de tanta diferença entre as médias de acertos. A principal diferença entre as três bases de teste é a origem dos dados. O *Twitter*® é um ambiente mais aberto, estando sujeito ao uso de uma linguagem mais livre, enquanto o *Youtube*® e o *Mercado Livre*® são ambientes nos quais os usuários estão ali especificamente para dar uma opinião e, portanto, utilizam a linguagem de forma mais consciente. Nos *websites* de *reviews* o usuário vai escrever algo para avaliar um objeto e, quase sempre, usa palavras de sentimentos extremos, ou seja, palavras que possuem um forte valor positivo ou negativo como, por exemplo, *excelente*, *muito bom*, *muito ruim*, etc.

Verificando os comentário do *Youtube*® e *Mercado Livre*® e a notória diferença nas taxas de acertos, pôde concluir, que a diferença é devido a linguagem utilizada.

Ao final dos testes pôde-se notar que realizar a análise de comentários do *Twitter*® é complicado, pois inúmeros fatores contribuem para um baixo rendimento na taxa de acerto. O *Twitter*® teve uma média de 46,4% de acertos comparando com as outras fontes avaliadas —*Youtube*® e *Mercado Livre*®, que obtiveram respectivamente 85,67% e 76%). Deste modo, pode-se provar que quando o ambiente é mais fechado a taxa de acerto sobe consideravelmente, pois neste caso os usuários se utilizam de termos especificamente de caráter opinativo para fazer o comentário. Porém, independentemente deste problema, o *Twitter*® se constituí uma excelente fonte de informações por estar em constante crescimento.

5 – CONSIDERAÇÕES FINAIS

A análise de sentimentos é uma área de pesquisa recente com poucos trabalhos desenvolvidos, principalmente para comentários em português. Tendo em vista tal cenário, e baseando-se no trabalho apresentado por (DENECKE, 2008), este trabalho apresentou um protótipo para classificar o sentimento dos comentários do *Twitter*® em português utilizando-se da *SentiWordNet*©.

DENECKE (2008) obteve para a língua inglesa e alemã, os valores de 51% a 62% e 59% a 66% de acertos respectivamente. Considerando tais resultados, inicialmente esperava-se que o protótipo desenvolvido atingisse tais marcas, o que não aconteceu, ficando com a média geral de 46,4% com desvio padrão de 12,38%. Após alguns levantamentos pôde-se observar que quando pesquisa-se algo no *Twitter*®, referente à uma entidade, muitos comentários considerados fatos são retornados, e estes não tem valor opinativo.

O trabalho de DENECKE (2008) foi realizado em cima de comentários de filmes. Para igualar os testes deste trabalho ao dele, foram coletados também comentários do *YouTube*® e *Mercado Livre*®, e pôde-se constatar que as taxas de acertos subiram consideravelmente. O fator relevante é que quando uma pessoa está em uma página de um produto os comentários realizados por este indivíduo serão relativos ao produto, diferente das redes sociais nas quais uma entidade é citada em um comentário sem que seja citado algo positivo ou negativo sobre ela.

Outro ponto a ser considerado na justificativa para o baixo índice de acertos é a excessiva utilização de abreviaturas, gírias e/ou termos informais por parte dos usuários do *Twitter*®, o que consequentemente contribui para uma baixa taxa de acerto. Mesmo utilizando um dicionário de abreviaturas este problema não foi completamente resolvido, pois a todo o momento novos termos são criados o que desatualiza o dicionário.

Existe também o problema da necessidade da tradução dos comentários, pois por vezes alguns termos que podem influenciar nos resultados não são traduzidos a

tradução. Além disso, neste trabalho não foi realizada a desambiguação dos termos devido ao custo computacional, o que deverá ser feito em trabalhos futuros.

Um ponto considerado relevante neste trabalho é que devido a técnica utilizada o protótipo realiza a análise de qualquer tipo de texto na língua portuguesa, indiferente do domínio de negócio. Apesar da taxa de acertos ser considerada baixa nos comentários do *Twitter*®, o protótipo respondeu bem na análise das opiniões do *YouTube*® e do *Mercado Livre*®.

Como trabalhos futuros sugerem-se a ampliação do pré-processamento do texto para distinguir os diferentes tipos de elementos sintáticos e verificar se existe alguma influência direta destes sobre o resultado das classificações; o projeto de um *webcrawler* específico para buscar comentários que possuem caráter opinativo, tentando identificar a entidade principal e extrair as características da entidade; a aplicação de técnicas de processamento de linguagem natural para tentar resolver os problemas de ambiguidade; e o desenvolvimento de um classificador que utilize a aprendizagem de máquina, com fácil adaptação a vários domínios de negócio.

REFERÊNCIAS

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F.. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Istituto di Scienza e Tecnologie dell‟Informazione, Consiglio Nazionale delle Ricerche , 2010.

DENECKE, K.. Using SentiWordNet for multilingual sentiment analysis.Data Engineering Workshop. ICDEW 2008. IEEE 24th International Conference on, 2008
.

G1. Globo. Brasil é o segundo país em número de usuários no Twitter, 2012. Disponível em: <http://g1.globo.com/tecnologia/noticia/2012/02/brasil-e-o-segundo-pais-em-numero-de-usuarios-no-twitter-diz-estudo.html>. Acesso em: 10-03-2012.

GUTEMBERG, N.. BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião. Monografia – Centro de Informática da Universidade Federal de Pernambuco. Recife,2010..

LIMA, D C. L. de A.. PairExtractor: Extração de Pares Livre de Domínio para Análise de Sentimentos. Monografia Centro de Informática da Universidade Federal de Pernambuco. Recife, 2011.

LIU. B.. Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data, Chicago, Spring, 2007. p. 479-514.

LOPES, T.. O que você acha da mineração?, 2009. Disponível em: <http://www.thlopes.com/2009/06/mineracao-texto-dados-opinioes/>. Acesso em: 15-03-2012.

MATIOLI, L.. Protótipo para mineração de opiniões em redes sociais: Estudo de casos selecionados usando o Twitter. Monografia Departamento de Ciência da Computação da Universidade Federal de Lavras. Minas Gerais, 2010.

OHANA, B. and TIERNEY, B.. Sentiment Classification of Reviews Using SentiWordNet. Dublin Institute of Technology, 9th. IT & T Conference, 2009.

PANG, B. e Lee, L. Opinion Mining and Sentiment Analysis, In Foundations and Trends in Information Retrieval, 2008.

SENTIMENT140. For Academics, 2012. Disponível em: <http://help.sentiment140.com/for-students>. Acesso em: 20-06-2012.

SILVA, N. R.; LIMA, D.; BARROS, F.. SAPair: Um Processo de Análise de Sentimento no Nível de Características. Brazilian Conference on Intelligent System, 2012.

SOLUCIONA. Equipe. E-commerce brasileiro fatura R$ 18,7 bilhões e cresce 26% em 2011, 2012. Disponível em: <http://www.comprafacilempresas.blog.br/e-commerce-2/e-commerce-brasileiro-fatura-r-187-bilhoes-e-cresce-26-em-2011/>. Acesso em: 10-03-2012.

SOUZA, L. V.. Análise de sentimentos no *Twitter* utilizando *SentiWordNet*. Monografia – Centro de Informática da Universidade Federal de Pernambuco. Recife, 2011.

WINTERWELL, Associates. JTwitter - the Java library for the Twitter API, 2011. Disponível em: <http://www.winterwell.com/software/jtwitter.php>. Acesso em: 23-04.2012.