GIULIA LUAN SANTOS DE SOUSA

TWEETMINING: ANÁLISE DE OPINIÃO CONTIDA EM TEXTOS EXTRAÍDOS DO *TWITTER*

LAVRAS – MG

2012

GIULIA LUAN SANTOS DE SOUSA

TWEETMINING: ANÁLISE DE OPINIÃO CONTIDA EM TEXTOS EXTRAÍDOS DO *TWITTER*

Monografia apresentada ao Colegiado do Curso de Sistemas de Informação, para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador

Prof. M.Sc. Eric Fernandes de Mello Araújo

LAVRAS – MG

2012

GIULIA LUAN SANTOS DE SOUSA

TWEETMINING: ANÁLISE DE OPINIÃO CONTIDA EM TEXTOS EXTRAÍDOS DO TWITTER

Monografia de graduação apresentada ao Colegiado do Curso de Sistemas de Informação, para obtenção do título de Bacharel em Sistemas de Informação.

APROVADA em 08 de outubro de 2012.

ANDRÉ GRUTZMAN

MARLUCE RODRIGUES PEREIRA

TIAGO AMADOR COELHO

ERIC FERNANDES DE MELLO ARAÚJO (orientador/a)

LAVRAS-MG

2012

Aos meus pais, Ana e Geraldo e aos meus irmãos, Douglas e Pedro.

AGRADECIMENTOS

Agradeço aos meus pais pela oportunidade que me deram para retribuir tudo o que me ensinaram, aos meus amigos que participaram desta jornada ao meu lado me dando o apoio necessário nos momentos que precisei e aos professores que se dedicaram a me transmitir conhecimento.

Agradeço, em especial:

A Rodolfo Barbosa, com quem luto junto e encontro apoio para alcançar um objetivo;

À Karen Bezerra, pessoa com quem dividi tantos momentos especiais nestes quatro anos;

A Eric Araújo, orientador que possibilitou a execução deste trabalho;

A Leandro Matioli, colega que sempre se dispôs a me auxiliar;

A Samuel Campos, Ronaldo Silva, Jaime Daniel Correa Mendes, Ivayr Farah Netto, Aleksander França e Luca Prieto pela primeira oportunidade e companheirismo na vida profissional;

À Alini Costa, Diuly e Diuliani Cristiani, Josiane Andrade, Tamara Amaral e Priscila Vaneli, a quem posso chamar de amigas;

Aos meus familiares, colegas e demais pessoas que aqui não mencionei mas que sabem da importância que representam em minha vida. Sinceros agradecimentos.

RESUMO

O objetivo do trabalho é implementar uma aplicação capaz de polarizar a opinião contida em textos extraídos do *Twitter* sobre a cidade de Campinas. Neste trabalho, utilizando o algoritmo *SVM* para mineração de opinião, os textos foram classificados em 2 classes inicialmente: neutros e opinativos, para que a partir dos documentos classificados, os *tweets* opinativos, ou seja, que expressam uma opinião do usuário a respeito do assunto abordado, fossem separados e montada uma nova base dividida entre textos contendo opiniões positivas e negativas. Assim, o sistema desenvolvido apresentou-se funcional, obtendo resultados satisfatórios para o problema proposto.

Palavras-Chave: Mineração de dados. Mineração de Opinião. Descoberta de Conhecimento em Base de dados. Mineração de Textos. Mineração Web. Aprendizado de Máquina. Processamento de Linguagem Natural.

ABSTRACT

The objective of this work is the implementation of an application capable of polarize the opinion contained in texts extracted from Twitter about the city of Campinas. Using SVM algorithms for opinion mining, at first, the texts were classified in two classes: neutral and opinative. After, from among the opinative, ie are able to express a opinion of user about a given matter, the tweets were classified on positive or negative opinions. This way the developed system has presented itself functional and achieved satisfactory results.

Keywords: Data Mining. Opinion Mining. Knowledge discovery in databases. Text Mining. Web Mining. Learning Machine. Natural Language Processing.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

1.1 Contextualização e Motivação

Com a crescente popularização das redes sociais e da Internet, tornou-se interessante para as grandes empresas buscar informações sobre seus produtos e serviços nesse meio, devido à significativa quantidade de dados relevantes encontrada nestas mídias. Neste cenário, considerando os avanços tecnológicos na área de mineração de dados que possibilitaram a extração dessas informações de maneira automatizada e viabilizaram a exploração destas informações, selecionou-se a cidade de Campinas-SP como objeto de estudo por vislumbrar a oportunidade de coletar uma grande quantidade de dados a respeito da cidade, pois esta oferece recursos suficientes para análise de polaridade que é o objetivo deste trabalho.

Considerando a velocidade de disseminação de textos na *web*, a necessidade da automatização do processo de busca e mineração destas informações é um fato recorrente uma vez que percebe-se a inviabilidade de executar este processo manualmente. Sendo assim, a falta de uma ferramenta rápida o bastante para buscar e processar estes dados pode trazer consequências que se agravam a cada minuto, pois o número de usuários que acessam e compartilham essa informação cresce exponencialmente.

Um exemplo simples da importância das informações compartilhadas nas redes sociais e do poder que elas podem exercer sobre a imagem de uma empresa em um curto período de tempo é o caso da Amazon.com, loja de vendas pela Internet, quando a empresa retirou alguns livros do ranking de vendas por uma falha de catalogação que levou dezenas de milhares de usuários de blogs e do *twitter* a acusar a empresa de censura, tudo isso em um período de pouco mais de 36 horas (TERRA, 2009).

Além da aplicabilidade nos negócios, é possível ainda extrair informações significantes nos meios sócio-cultural e políticos de uma região, as opiniões dos usuários sobre política e governo, cultura e outros aspectos de um local. Tendo em mãos estas informações, torna-se oportuna a implementação de melhorias no local estudado, pesquisas de satisfação dos moradores, melhores aplicações para investimentos, dentre outras possibilidades.

Espera-se que os resultados levantados nesse trabalho sejam de grande importância e permitam mapear a maneira como os usuários das redes sociais analisadas avaliam, negativamente ou não, os assuntos em questão, podendo auxiliar nas tomadas de decisões, na tentativa de potencializar os ganhos e diminuir os custos com eventuais problemas de ineficiência dos investimentos.

1.2 Objetivos do Trabalho

Nesse contexto, objetiva-se o desenvolvimento de uma ferramenta que seja capaz de extrair informações sobre os sentimentos dos usuários da rede social *Twitter* sobre a cidade de Campinas, com o intuito de montar um mapeamento destes sentimentos e permitir a análise dos mesmos através da polarização dos *tweets*.

Os seguintes objetivos específicos foram alcançados:

- Realização de um levantamento bibliográfico dos métodos utilizados no trabalho, apresentando os conceitos necessários para o entendimento do mesmo;
- Estudo dos algoritmos de mineração de opinião e análise de sentimentos e sua aplicação para encontrar padrões e características relevantes sobre a cidade de Campinas-SP;
- Implementação da aplicação de mineração dos dados extraídos.

1.3 Trabalhos Relacionados

Este trabalho busca, por meio da mineração de opinião, analisar aspectos no contexto social e geopolítico da cidade de Campinas, visto que a maior parte da pesquisa atual na área de mineração de opinião tem se concentrado em produtos e negócios, o que leva a detecção de uma deficiência nesta área a ser pesquisada.

A partir do levantamento acerca das pesquisas na área de mineração de opinião em mídias sociais, foram encontrados alguns trabalhos relacionados. Os principais estão listados abaixo:

- *Twitter Sentiment*: Trata-se de uma ferramenta que permite ao usuário inserir uma palavra-chave para análise utilizando a base de dados do *Twitter*. A partir da pesquisa realizada, a aplicação apresenta gráficos de análise de polaridade e as postagens utilizadas na análise, para que o usuário analise a acurácia da ferramenta[1];

- Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o *Twitter*: Foi desenvolvido um protótipo para análise de polaridade utilizando postagens de usuários do *Twitter* sobre o *Windows 7*, a empresa *Apple* e o filme *Jackass 3D*(MATIOLI, 2010);

- *SentimineTM*: Sistema desenvolvido pelo Grupo *Parnassus* que permite a comparação de diferentes produtos e a análise de polaridade de produtos utilizando postagens de blogs[2];

- *Salience*: Motor de análise de textos multi-lingual desenvolvido pela *Lexalytics* integrado a sistemas de inteligência de negócios, monitoramento de

[1] A aplicação pode ser visualizada em http://twittersentiment.appspot.com/

[2] O site oficial do projeto é http://sentimine.com/

mídias sociais, automatização de negociações, análise de pesquisas, entre outras[3].

1.4 Organização do Trabalho

O trabalho está estruturado como se segue:

No capítulo 1 vimos uma introdução ao trabalho, contento uma breve contextualização, motivação, os objetivos gerais e específicos, a metodologia adotada, trabalhos relacionados e o cronograma do trabalho.

No capítulo 2 é apresentada uma revisão dos conceitos fundamentais de descoberta de conhecimento em base de dados, mineração de dados, mineração de textos, mineração na *web*¸ mineração de opinião e análise de sentimentos, aprendizado de máquina e extração de dados em redes sociais.

O capítulo 3 apresenta a metodologia utilizada.

O capítulo 3 mostra o processo de implementação do sistema e comentários sobre sua construção.

O capítulo 4 contém os resultados obtidos com a implementação do trabalho e discussões acerca dos objetivos alcançados.

O capítulo 5 discorre sobre as conclusões acerca da validade do trabalho realizado e recomendações para trabalhos futuros, visando continuar a linha de pesquisa desenvolvida pelo mesmo.

[3]Mais informações sobre o projeto podem ser visualizadas em http://www.lexalytics.com/

2 REFERENCIAL TEÓRICO

Para compreensão e entendimento desse trabalho, é importante conhecer conceitos de análise de sentimentos e mineração de opinião, subáreas de mineração de dados. São importantes também os conceitos sobre processamento de linguagem natural, aprendizagem de máquina e alguns dos algoritmos mais relevantes no processo de extração de opinião a partir de textos que serão apresentados adiante.

Apresenta-se ainda uma contextualização das redes sociais, que são a fonte de dados deste trabalho.

2.1 Descoberta do conhecimento em bases de dados (*KDD*)

No final da década de 80 surgiu uma nova área de pesquisa que visa extrair conhecimento de bases de dados de maneira computacional através do uso de ferramentas de *KDD* (*Knowledge Discovery in Databases*) ou descoberta do conhecimento em bases de dados.

(FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1997) conceituam KDD como uma maneira não trivial de identificar padrões válidos que inicialmente são desconhecidos porém potencialmente úteis a partir de uma base de dados. (GRAY; WATSON, 1999) reforçam este conceito ao dizer que *KDD* foi projetado para extrair informações gerenciais importantes aos gestores que não podem ser reveladas a partir de consultas e relatórios de maneira eficaz.

"Segundo (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992), dado um conjunto de fatos (dados) F, uma linguagem L, e alguma medida de certeza C, foi definido um padrão como uma declaração S em L, que descreve relações entre um subconjunto FS de F com uma certeza C, tal que S é mais simples (em algum sentido) do que a enumeração de todos os fatos em FS. Um padrão que é

interessante (de acordo com uma medida de interesse do usuário) e determinante o suficiente (mais uma vez de acordo com critérios do usuário) é chamado de conhecimento. A saída de um programa que monitora o conjunto de fatos em um banco de dados e produz padrões neste sentido, é a descoberta de conhecimento."

Alguns autores afirmam que o fator maximizante do desempenho das operações de mineração em grandes conjuntos de dados persistentes é o componente de banco de dados do *KDD* e que a descoberta de conhecimento é simplesmente a aprendizadem de máquina a partir desses conjuntos. Deve-se reconhecer a importância do componente de banco de dados do *KDD* e a melhora do desempenho do mesmo, porém individualmente ele provavelmente não é suficiente para desencadear uma mudança qualitativa nas capacidades de um sistema (IMIELINSKI; MANNILA, 1996).

Devido ao grande fluxo de dados que tem sobrecarregado as comunidades corporativas, governamentais e científicas, a exploração de ferramentas de *KDD* se torna uma área de pesquisa promissora, pois o volume de dados é grande e a análise destes dados com o objetivo de extrair padrões significativos em tempo hábil é um problema intratável sem ajuda computacional. Porém sem a ajuda de humanos para interpretar corretamente os resultados, *KDD* se torna desnecessário, pois o grande desafio encontrado na descoberta do conhecimento em bases de dados é processar uma grande quantidade de dados brutos e apresentá-los ao usuário em forma de conhecimento, mostrando os padrões mais importantes e significativos detectados no processamento (MATHEUS; CHAN; PIATETSKY-SHAPIRO, 1993). Assim, *KDD* apresenta-se como uma potencial ferramenta na tentativa de resolver o problema da sobrecarga de dados no meio digital.

Diversos periódicos científicos de relevância[1] apresentaram, nos últimos anos, trabalhos em torno de aplicações de *KDD*, o que aumenta o interesse na área e pode

[1] *Business Week, Newsweek, Byte, PC Week* e outros

alavancar uma nova geração de ferramentas e teorias computacionais que auxiliem na extração de informação a partir de dados brutos, uma vez que o volume de dados cresce rapidamente (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1997).

(IMIELINSKI; MANNILA, 1996; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1997) dividem o processo de *KDD* em cinco etapas. São elas: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação do conhecimento extraído. Essas fases e suas interações podem ser vistas de maneira geral na figura 2.1. (BRACHMAN; ANAND, 1996) ressaltam a importância da interação humana nas etapas do processo devido a importância da orientação na execução do processo por um agente que possua conhecimento sobre o domínio do problema tratado.

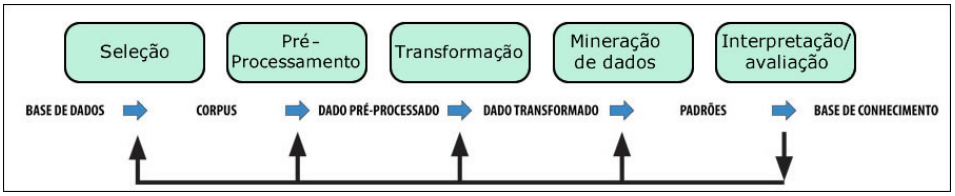

Figura 2.1: Etapas que constituem o processo de *KDD*. Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1997).

Como se pode verificar, *data mining* ou mineração de dados, é uma das etapas do processo de *KDD*, talvez a principal, uma vez que é nessa etapa que as informações são extraídas de fato. Um maior detalhamento da mesma será apresentado na seção 2.2. Porém, uma atenção especial deve ser dada à fase de pré-processamento dos dados, que tenta identificar e remover os problemas que comumente se apresentam nos dados extraídos de bases reais, tais como: grande quantidade de ruído e inconsistências, desproporção entre as distribuições das classes (classes desbalanceadas), excesso de valores desconhecidos, entre outros, pois a qualidade dos dados fornecidos como entrada está diretamente relacionada à qualidade do conhecimento extraído no processo (BATISTA, 2003).

(BATISTA, 2003) propõe duas classificações para as tarefas realizadas na fase de pré-processamento dos dados: tarefas fortemente dependentes de conhecimento de domínio e tarefas fracamente dependentes de conhecimento de domínio. As tarefas fortemente dependentes de domínio são aquelas que somente podem ser efetivamente realizadas com o uso de conhecimento específico do domínio tratado. Já nas tarefas fracamente dependentes de conhecimento de domínio, as informações necessárias para tratar os problemas de pré-processamento dos dados podem ser extraídas dos próprios dados, apresentando um maior grau de automação.

2.2 Mineração de dados

Nesta seção serão apresentados os principais conceitos acerca de mineração de dados e as diferentes formas de mineração.

(GRAY; WATSON, 1999) subdividem a mineração de dados de acordo com os tipos de dados e objetivos em cinco grupos: associação, sequência, classificação, clusterização e previsão. As associações procuram descobrir relações em determinados conjuntos de dados. Por exemplo, a compra de pães associa-se à compra de itens como margarina, requeijão, entre outros, que em alguns casos podem parecer óbvios, mas em outros nem tanto. As sequências identificam acontecimentos que levam a outros. Por exemplo, a compra de uma casa leva a compra de móveis. Classificações e clusterizações têm objetivos parecidos, pois dividem os dados em grupos com indivíduos semelhantes através do reconhecimento de padrões entre eles, porém elas se diferem no ponto em que a classificação é feita de acordo com amostras das classes oferecidas *a priori*, enquanto a clusterização é feita a partir de dados desconhecidos introduzidos ao algoritmo a fim de que o mesmo reconheça padrões e extraia informações suficientes para separar grupos de dados (*clusters*). As previsões analisam uma série temporal e preveem os dados futuros de acordo com as informações extraídas da série.

(KANTARDZIC, 2011) adiciona algumas categorias a esta subdivisão como sumarização, que envolve métodos para resumir um conjunto de dados em uma descrição mais compacta, e detecção de desvios, que tenta descobrir mudanças significativas nos padrões dos dados a fim de identificar alguma irregularidade em uma série de dados estudada.

(BERRY; LINOFF, 2004) defendem a importância da mineração de dados no mundo dos negócios quando diz que as decisões tomadas a partir de informações que justifiquem a escolha são melhores que aquelas tomadas arbitrariamente. Também acrescentam que a mineração de dados se apresenta como uma ferramenta para conseguir informações que embasem essas decisões. Frequentemente pode-se perceber o uso de mineração de dados nesse contexto. Por exemplo, a *Walmart*, multinacional americana de lojas de departamento, desde o final dos anos 90 adota um sistema de pesquisa para prever a demanda de cada item do estoque analisando os padrões de compras dos clientes e descobrindo vínculos entre os itens comprados. O *Bank of America* utiliza técnicas de mineração de dados para selecionar clientes com menor risco de inadimplência e traçar estratégias de marketing focadas nestes clientes visando o lucro da organização.

Assim, mineração de dados em suas diversas classificações, pode ser usada para descoberta de conhecimento em diversas áreas, auxiliando nas pesquisas científicas e mercadológicas e se mostrando como uma ferramenta poderosa para obtenção de resultados satisfatórios nos diversos ramos de aplicação possíveis (BASTOS, 2001; MONTINI, 2009).

2.2.1 Mineração de Textos

A mineração de textos tenta recolher informações significativas a partir de uma grande quantidade de textos escritos em linguagem natural.

Mineração de texto pode também ser definido como a aplicação de algoritmos e métodos de apredizagem de máquina e estatísticos para textos com o objetivo de encontrar padrões úteis. Para este propósito, é necessário para pré-processar os textos em conformidade (HOTHO; N'RNBERGER; PAAB, 2005).

Os textos geralmente apresentam mais dificuldade na extração de informações quando comparados a dados armazenados em bancos de dados, por estarem mais desestruturados e mais difíceis de serem pré processados para a inserção nos algoritmos de mineração. Na cultura moderna, é necessário trabalhar com dados em forma de textos, já que este é o veículo comumente utilizado para o intercâmbio de informações, o que torna este campo de pesquisa bastante interessante (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1997).

A grande diferença entre mineração de dados e mineração de textos não é apenas a fonte de informação, mas o fato de que em mineração de dados, a informação extraída está implícita e é desconhecida (WITTEN; FRANK, 2000), já na mineração de textos, a informação a ser extraída está explicitada no texto a ser minerado, porém com a grande quantidade de textos torna-se inviável minerá-los de maneira não computacional.

A mineração de textos pode ser dividida em diversas etapas. (DIJRRE; GERSTL; R., 1999) dividem em três etapas, que consistem na identificação de um conjunto, preparação e seleção de recursos e análise da distribuição. Já (MATHIAK; ECKSTEIN, 2004), dividem-na em cinco etapas, que são: coleta dos dados; pré-processamento; análise; visualização; e evolução. Como ainda não existe um consenso na literatura para esta divisão, neste trabalho será utilizada a abordagem proposta por (ARANHA; VELLASCO, 2007) que propõe cinco etapas mostradas na figura 2.2.

A etapa da coleta consiste na obtenção dos textos que serão utilizados no processo de mineração formando a base de documentos denominada *Corpus*. A

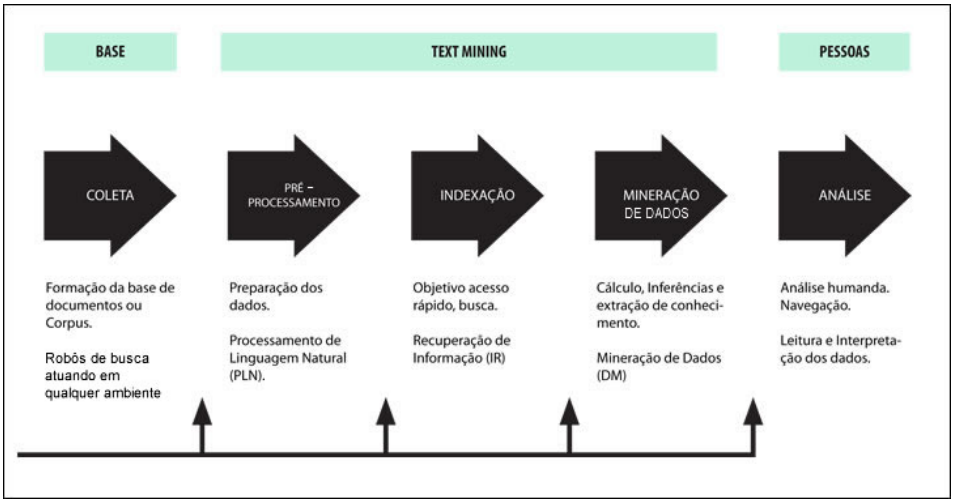

Figura 2.2: Etapas que constituem o processo de Mineração de Textos.Adaptado de (ARANHA; VELLASCO, 2007)

etapa de pré-processamento prepara os dados para serem trabalhados computacionalmente. É uma fase importante, uma vez que a qualidade dos resultados é diretamente influenciada pela qualidade dos dados de entrada, sendo assim, se estes não forem pré-processados adequadamente poderão prejudicar os resultados. Essa fase é também bastante custosa devido à natureza dos dados.

Em terceiro vem a etapa de indexação que tem como objetivo categorizar os dados dos documentos de maneira que facilite a busca e agilize o acesso aos dados.

A quarta etapa, que é a mineração de dados em si, é executada através de técnicas de cálculos e inferências com o intuito de extrair informação dos dados.

Na quinta etapa, inicia-se a análise dos resultados, que conta com a interferência humana para interpretá-los e utilizá-los da melhor maneira possível.

Este processo pode ainda ser realimentado em qualquer uma de suas etapas caso seja verificada a necessidade de voltar a uma etapa para obtenção de melhores resultados.

2.2.2 Mineração na *web*

O número de usuários de internet no mundo tem aumentado consideravelmente com o passar dos anos. Em 1995, existiam aproximadamente 45,1 milhões de usuários, em 2000 esse número cresceu para 420 milhões, em 2005 ultrapassou-se a barreira dos bilhões com 1,08 bilhões de usuários e em setembro de 2009 já existiam 1,73 bilhões (CURY, 2010). Esse número não parou de crescer e em 30 de junho de 2012, existiam 2.405.510.175 usuários segundo o site internetworldstats.com (GROUP, 2012). Esse constante aumento de usuários fez com que atualmente, a maior parte das aplicações e informações estejam disponíveis na *web*, assim, a tentativa de extração de conhecimento nesse meio tornou-se bastante interessante já que estas informações têm acesso público e um conjunto de dados bastante rico. A tentativa de extração de conhecimento baseada na mineração dos dados encontrados na web deu-se o nome de *web mining*, ou mineração na *web* (KOSALA; BLOCKEEL, 2000; ETZIONI, 1996).

O processo de *web mining* pode ser dividido em três categorias que se subdividem em outras como se pode visualizar na figura 2.3:

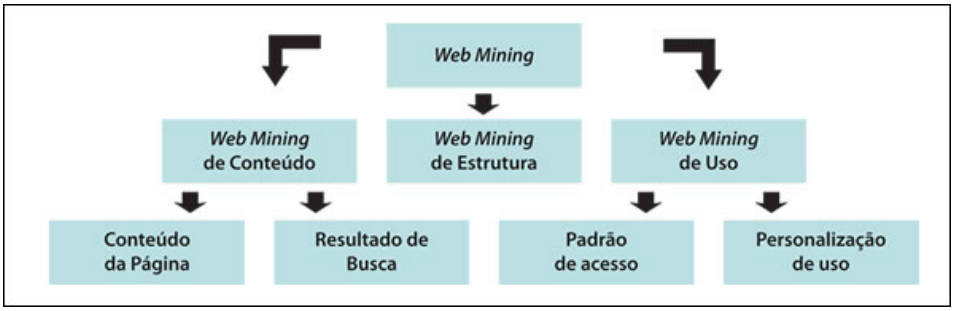

Figura 2.3: Estrutura da mineração na *web*. Adaptado de (JUNIOR; PASSOS, 2007)

A *web mining* de conteúdo, que é o foco desse trabalho, consiste em minerar o conteúdo extraído de textos, imagens, áudios, vídeos e registros estruturados como

listas e tabelas. Grandes empresas do ramo de buscas, como o *Google*[2] , executam esse tipo de mineração a fim de relacionar os parâmetros de busca dos usuários com o conteúdo extraído e tratado (SRIVASTAVA; DESIKAN; KUMAR, 2002).

Web Mining de estrutura consiste na tentativa de extração de informação baseada nas ligações entre as páginas através dos *hiperlinks* contidos nas mesmas, podendo-se verificar a popularidade de uma página em um determinado contexto através da quantidade de referências dela por outras páginas e outras características que podem ser extraídas a partir da análise da estrutura de grafos da *web*, como o rankeamento de páginas feito por aplicações como o *PageRank* [3] (CROFT; METZLER; STRHOMAN, 2009; SINGH; KUMAR, 2009; SHARMA; TYAGI; BHADANA, 2010) e o *Hyperlink-Induced Topic Search (HITS)* [4] (JAIN; PUROHIT, 2011).

Por último tem-se a *web mining* de uso, que visa utilizar técnicas de mineração para descobrir padrões de uso das aplicações na tentativa de prever o comportamento do usuário ao interagir com a *web* a partir dos dados gerados pelas transações cliente-servidor. Estas informações podem ser aplicáveis desde a sugestão de melhorias na estrutura de um site, como o posicionamento de *hiperlinks* e outras coisas, a indicações de produtos ao usuário baseado na frequência em que o mesmo realiza compras pela internet e nos produtos que ele compra, por exemplo (YI-XING; WEI; ZHEN-HUA, 2010).

Como dito anteriormente, este trabalho se situa na área categorizada como mineração de conteúdo da *web*, uma vez que será analisado o conteúdo encontrado usando o serviço *Twitter* como base de dados.

[2] http://www.google.com.br

[3] *PageRank* é a forma pela qual o *Google* procura representar a importância que um site, ou página, tem para ele (*Google*) frente a *Internet*. Ele foi desenvolvido em 1995 na Universidade de Stanford por Larry Page, daí vem o nome *Page Rank*

[4] É o precursor do *PageRank*, desenvolvido por Jon Kleinberg

2.3 Mineração de Opinião

A seção que segue mostra os conceitos necessários ao entendimento do significado de mineração de opinião trazidos por pesquisadores reconhecidos na área, possibilitando assim a compreensão deste trabalho.

Aborda-se conceitos de análise de sentimento, o principal ramo da mineração de opinião e processamento de linguagem natural.

O termo mineração de opinião surgiu oficialmente em 2003, sendo definido como uma ferramenta capaz de agregar opiniões sobre os atributos de um item a partir do processamento dos resultados de uma pesquisa sobre o mesmo (PANG; LEE, 2008; DAVE; LAWRENCE; PENNOCK, 2003).

A partir daí, o termo, já popular nas comunidades ligadas a análise de conteúdo na *web* e recuperação de informação, ganhou interpretações mais amplas.

(CHEN; ZIMBRA, 2010) definem mineração de opinião como uma subárea da mineração de dados que se refere a técnicas para extração, classificação, processamento e entendimento de opiniões expressas em diversas fontes de informação *on line*, comentários em redes sociais e outros mecanismos usados por indivíduos para expressar suas opiniões.

As mídias *online* podem conter opiniões interessantes sobre produtos ou ainda a respeito de aspectos geopolíticos e sociais, o que permite extrair informações e sentimentos baseados no aspecto subjetivo contido nos textos e classificá-las como positivas ou negativas de acordo com um aspecto estabelecido (CHEN; ZIMBRA, 2010). Porém o grande problema da mineração de opinião é a complexidade encontrada em sentimentos expressos por um grande grupo de participantes, uma vez que eles por si só não permitem aos pesquisadores total entendimento das opiniões

expressas pelos mesmos, necessitando de uma análise mais aprofundada acerca dos fatos para conceituar uma opinião positiva ou negativa.

As etapas da mineração de opinião são as mesmas representadas na figura 2.2.

2.3.1 Processamento de Linguagem Natural

O que distingue as aplicações de processamento de linguagem de outros tipos de aplicações de processamento de dados é o uso do conhecimento da linguagem (JURAFSKY; MARTIN, 2009).

Sistemas que usam do processamento de linguagem natural para realizar suas tarefas necessitam de conhecimento da linguagem trabalhada. (JURAFSKY; MARTIN, 2009) citam os tipos de conhecimento necessários para se entender os comportamentos da linguagem natural listando-os em seis tipos:

1. Fonética e fonologia: conhecimento sobre sons linguísticos;
2. Morfologia: conhecimento dos componentes significativos das palavras, estudando sua estrutura e analisando-as isoladamente dentro de um período;
3. Sintaxe: conhecimento das relações estruturais entre as palavras;
4. Semântica: conhecimento do significado das palavras;
5. Pragmática: conhecimento da relação de significado para os objetivos e intenções do locutor;
6. Discurso: conhecimento sobre as unidades linguísticas maiores que um discurso único.

Por exemplo, sistemas de reconhecimento de fala requerem conhecimento sobre fonética e fonologia enquanto sistemas de tradução precisam conhecer desde a

morfologia e sintaxe das palavras às locuções e contrações que as mesmas podem apresentar em diferentes contextos em que forem empregadas. Outro aspecto que deve ser considerado nestes sistemas são as fraseologias da língua, que são expressões frequentemente utilizadas pelos indivíduos como “dar com os burros n’água” ou “pau pra toda obra”, que podem trazer um significado importante na análise da linguagem.

Em suma, o objetivo dos sistemas de processamento de linguagem natural é tornar os computadores capazes de realizar tarefas úteis que envolvem a linguagem humana, permitindo uma melhora na comunicação homem-máquina ou simplesmente fazendo processamento de textos de maneira útil e gerando valor para o usuário (JURAFSKY; MARTIN, 2009).

2.3.2 Análise de Sentimentos

Segundo (CHEN; ZIMBRA, 2010) a análise de sentimento é uma ferramenta utilizada em mineração de opinião com o objetivo de identificar sentimentos expressos pelo usuário em seus textos. (LIU, 2006) define como um estudo computacional de opiniões pessoais, avaliações, emoções e outros aspectos que o usuário pode inserir em seus textos.

A análise de sentimentos é um campo em crescente expansão, uma vez que os usuários das redes sociais e outras mídias como *blogs*, estão cada vez mais opinando sobre os mais diversos assuntos nestas mídias, gerando assim um grande interesse na área.

Grandes organizações como Walmart, MCDonalds, CIA, entre outras, estão investindo em grupos de pesquisa em análise de sentimentos e mineração de opinião, pois percebem o valor das opiniões expressas nas mídias sociais e o quanto estas opiniões podem afetá-las de maneira positiva e negativa (CHEN; ZIMBRA, 2010). (LIU, 2006) ressalta que as pessoas costumam se importar com outras opi-

niões ao tomar suas decisões. Outros fatores motivam as grandes corporações a investirem na área de extração e análise de sentimentos, como a influência das mídias sociais nos preços das ações, as tendências para criação de novos produtos e melhorias nos já existentes. A direção para novas estratégias de negócio e muitas outras vertentes que a extração de conteúdo do usuário e os sentimentos expressos por ele podem atingir. Outro fator interessante e motivador é que essa abordagem pode ser utilizada em todos os contextos onde as opiniões podem ser mineradas e classificadas a partir de uma grande quantidade de textos (ESULI; SEBASTIANI, 2010).

Porém a análise de sentimentos em mídias sociais não é um problema simples de ser tratado, pois é constituído por uma combinação de diversos subproblemas, faces e em muitos casos as opiniões estão escondidas em *posts* com diversas respostas (ABASSI, 2010), o que torna inviável o monitoramento por um ser humano, uma vez que este trabalho tomaria um longo tempo ou uma grande quantidade de pessoas dedicadas a esta tarefa, tornando o custo muito alto. A seguir são expostas algumas dificuldades mais comuns encontradas no processo de análise de sentimentos.

2.3.3 Principais dificuldades na análise de sentimentos e mineração de opinião e alternativas propostas

As dificuldades encontradas na análise de sentimentos em textos na web são muitas, visto que computadores são excelentes máquinas de calcular números. Porém, calcular sentimentos é uma tarefa difícil até mesmo para seres humanos, pois a análise de aspectos como ambiguidade, ironia, sarcasmo, gírias e outras formas de expressão presentes na linguagem natural devem ser considerados.

Um dos primeiros problemas identificados na extração e análise de sentimentos é filtrar os textos que realmente expressam alguma opinião acerca do assunto

pesquisado, visto que a maioria das postagens dos usuários não são opinativas. Isso pode ser comprovado analisando os resultados de diversos softwares desenvolvidos para análise de polaridade, em que as opiniões são classificadas entre neutras ou opinativas e posteriormente, dentre as opinativas, agrupadas em positivas e negativas. O software desenvolvido pela *Sentimine*, de Seatle, que observou a avaliação do *Kindle DX*, um leitor de *ebooks* produzido pela Amazon, verificou que de 1500 mensagens analisadas, 89% eram neutras, ou seja, não expressavam nenhuma opinião positiva ou negativa, o que dificulta a obtenção de uma base de dados relevante para a análise (SIMONITE, 2009). Uma alternativa a esse problema é remover os textos que não expressam opinião e a partir daí, fazer uma análise mais crítica acerca dos textos opinativos.

Outro empecilho levantado é a forma como estes textos são encontrados, pois as mídias sociais não possuem um padrão de postagem e os formatos variam de uma mídia para outra. Problemas originados pela maneira como o usuário expressa suas opiniões, o uso de abreviaturas como “vc”, “fds” para dizer você e fim de semana, respectivamente, textos escritos com palavras erradas e com sentenças sintaticamente mal formadas, o uso de pronomes como “ele” para referenciar produtos ou a referência a mais de um item numa mesma sentença são comumente encontrados nas postagens dos usuários e podem dificultar o processamento dos textos. Para amenizar este problema, pode ser criado um dicionário de termos, que possibilita que estas folksonomias sejam substituídas por taxonomias.

Os resultados desejados também podem significar um problema. Pode-se esperar extrair aspectos referentes a características do item observado, desejando-se obter dados sobre os quesitos que mais necessitam de melhorias, ou analisar divergências entre as opiniões dos usuários, que envolve muito mais dificuldade que uma categorização entre opiniões positivas e negativas. Assim, deve-se analisar o custo-benefício de cada uma destas abordagens para tomar a decisão sobre qual delas utilizar.

Por fim, um último problema é a representação dos resultados de maneira que implique numa fácil compreensão e permita a extração do maior número de informações possíveis a partir dos mesmos. Duas alternativas são a sumarização dos resultados em forma textual e gráfica, sendo a primeira bem menos trivial, pois deve-se considerar a geração de linguagem natural com base nos conteúdos, enquanto a representação gráfica, por sua vez é mais simples, porém a escolha errada das informações presentes no gráfico pode acarretar numa difícil compreensão e interpretação dos resultados.

2.3.4 Aplicações

Muitas aplicações já estão sendo implementadas nesta área, desde as mais comuns, que avaliam polaridade de sentimentos, à aplicações que tentam prever tendências e até mesmo influenciar as opiniões dos usuários.

Softwares de inteligência de negócio são as aplicações mais comuns quando se fala em mineração de opinião e análise de sentimentos. Sistemas de classificação de polaridade de opiniões e até mesmo aqueles que conseguem detectar aspectos positivos e negativos de um produto são úteis às organizações (IBM, 2012b).

Sistemas que analisam aspectos governamentais também são úteis para monitorar opiniões tanto no que se refere à aspectos geopolíticos de uma região quanto em pesquisas sobre eleições e outros acontecimentos que interferem no poder administrativo de uma cidade ou até mesmo um país (CATE, 2008).

Dentre as diversas aplicações possíveis, a mineração de opinião também pode ser utilizada como complemento de outros aplicativos, como *softwares* de recomendação, analisando os produtos a serem recomendados e excluindo aqueles que receberem uma avaliação negativa dos usuários ou servindo de base para uma ordenação dos que possuem uma avaliação significativa (IBM, 2012a) .

2.4 Aprendizagem de Máquinas

Aprendizagem de máquina é uma subárea de pesquisa em inteligência artificial que estuda métodos computacionais para adquirir comportamento inteligente e simular o aprendizado humano a partir de computadores por meio de indução (MICHALSKI; CARBONELL; MITCHEL, 1983). A seguir são apresentados alguns conceitos encontrados na literatura sobre esta subárea.

"(SIMON, 1984) define aprendizado como qualquer mudança num sistema que melhore o seu desempenho na segunda vez que ele repetir a mesma tarefa, ou outra tarefa da mesma população".

"(BISHOP, 2003) conceitua como uma área de Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o processo de aprendizado".

Em geral, o aprendizado de máquina é subdividido em supervisionado, semi-supervisionado e não supervisionado.

No aprendizado supervisionado, que é o método utilizado neste trabalho, um modelo de classificação é construído utilizando um conjunto de instâncias que representam cada classe para treinamento. A partir destes conjuntos o modelo é capaz de prever associações de classe para novas instâncias desconhecidas examinando suas características. Para classificação das instâncias de treinamento é necessária intervenção de um agente especialista (WILLIAMS; ZANDER; ARMITAGE, 2006).

No aprendizado semi-supervisionado, assim como no supervisionado, um modelo é construído a partir de instâncias previamente classificadas por um especialista complementadas por instâncias não classificadas e a partir deste modelo

realiza-se o treinamento do algoritmo caracterizando o aprendizado (CIRELO; COZMAN, 2003).

O aprendizado não supervisionado é caracterizado pela incerteza sobre a saída esperada, ou seja, não se tem conhecimento *a priori* sobre a classificação das instâncias e o agrupamento é feito baseado em fatores comuns identificados pelo algoritmo. Este tipo de aprendizado é comumente utilizado para resolver problemas de clusterização, sobre os quais não se conhece o número de *clusters* ou características em comum entre as instâncias.

2.4.1 Support Vector Machines (*SVMs*)

As *SVMs* surgiram no final dos anos 70 (VAPNIK, 1982) e desde então têm sido cada vez mais utilizadas para resolver problemas de diversas naturezas como reconhecimento de dígitos manuscritos, reconhecimento de sons e imagens, estimativa de densidade (WESTON *et al.*, 1998), classificação de textos, dentre outras aplicações.

Elas são classificadas em dois tipos, *SVMs* lineares, que definem fronteiras lineares para a separação de dados pertencentes a diferentes classes (LORENA; CARVALHO, 2007), e *SVMs* não lineares utilizadas em casos em que não se pode separar os dados por um hiperplano num plano X,Y de duas dimensões.

2.4.1.1 SVMs lineares

As *SVMs* lineares definem a partir de um conjunto de treinamento, um conjunto de hiperplanos que dividem um espaço de dados em regiões que representam suas classes.

As *SVMs* de margens rígidas tratam de problemas linearmente separáveis por hiperplanos de maneira que não haja dados de treinamento entre as margens de

separação das classes como se pode ver na figura 2.4, a partir dos hiperplanos que separam as classes é considerado aquele com melhor capacidade de generalização, ou seja, aquele que é capaz de separar as classes com maior margem entre o hiperplano e as instâncias de cada classe mais próximas ao mesmo (LORENA; CARVALHO, 2007).

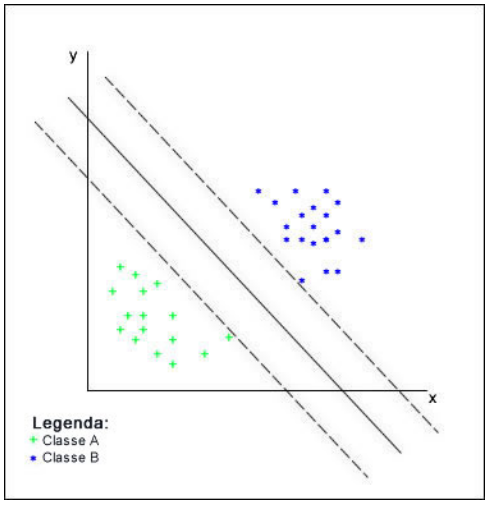

Figura 2.4: Separação das classes A (+) e B (*) por um hiperplano linear no espaço x,y

Porém, em situações reais, as aplicações em sua maioria possuem dados que não podem ser linearmente separáveis sem nenhum erro. Para tratar estes problemas foram criadas as *SVMs* de margens suaves, que permitem um pequeno erro em sua classificação (LORENA; CARVALHO, 2007), ou seja, algumas instâncias de uma determinada classe estar localizada em uma região de fronteira do hiperplano ou em uma região diferente da que a representa no espaço, como se pode verificar na figura 2.5.

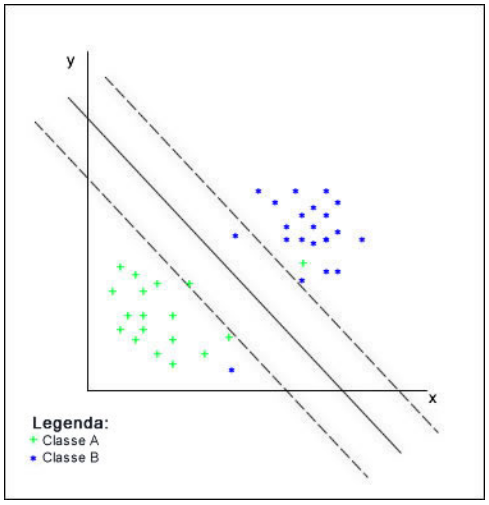

Figura 2.5: Separação das classes A (+) e B (*) por um hiperplano linear no espaço x,y com ruído

2.4.1.2 SVMs não lineares

Alguns problemas possuem conjuntos de dados que não podem ser linearmente separados de maneira satisfatória, dessa maneira, o uso de uma fronteira curva seria mais adequado para separação das classes, como mostra a figura 2.6.

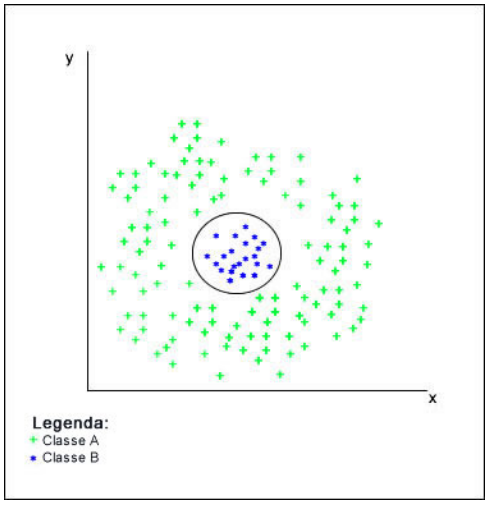

Figura 2.6: Representação de classes não linearmente separáveis

Para o tratamento destes problemas, foi criada uma extensão às *SVMs* denominada *SVMs* não lineares (LORENA; CARVALHO, 2007), que mapeiam o conjunto de treinamento de seu espaço original para um novo espaço de maior dimensão, chamado espaço de características, onde estes dados possam ser linearmente separáveis, tornando possível o uso das *SVMs* para a resolução dos mesmos. Para que este procedimento seja executado, duas condições devem ser satisfeitas, a primeira é que a transformação seja não linear e a segunda é que a dimensão do espaço de características seja suficientemente alta para separar as classes. Estes conceitos são ilustrados na figura 2.7.

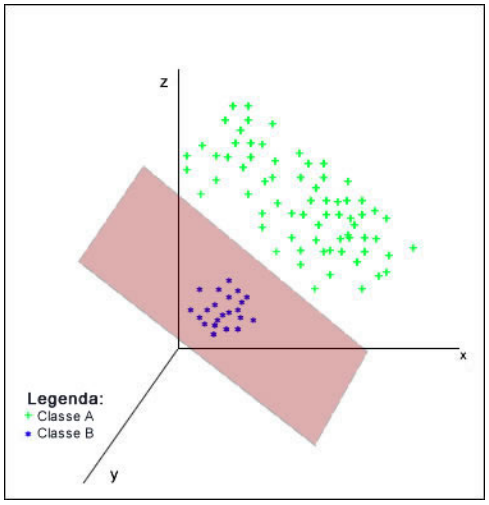

Figura 2.7: Representação de classes não linearmente separáveis mapeadas para um espaço de maior dimensão onde as classes são linearmente separáveis

2.4.2 Outros algoritmos

Diversos outros algoritmos podem ser utilizados para extração e análise de opinião a partir de textos. Os algoritmos bio inspirados (*bio-inspired algorithms - BIA*) são algoritmos poderosos que tentam imitar o comportamento de entidades biológicas de maneira computacional e pode ser aplicado ao problema de classificação de textos (BRABAZON; O'NEILL, 2010).

Vários trabalhos realizados na área fazem uso de algoritmos baseados no comportamento de enxames (PRIOR; CASTRO, 2010) e formigueiros. Existem também algoritmos conhecidos como o *K-Means* (WAGSTAFF *et al.*, 2001), *DBSCAN* (MOREIRA; SANTOS; CARNEIRO, 2005), *CLONALG* (BROWNLEE, 2005), dentre outros que não são objetos de estudo deste trabalho.

Outros algoritmos que vem sendo utilizados com este propósito são os algoritmos baseados em redes neurais artificiais (RNAs). "Segundo (MEHRA; WAH, 1992), RNAs são estruturas computacionais modeladas em processos biológicos", porém, apesar de as redes neurais artificiais terem sido inspiradas pelas redes de neurônios biológicos, as semelhanças foram diminuindo ao longo dos anos com a evolução das pesquisas na área (BARRETO, 2002). Uma outra definição dada por "(BRABAZON; O'NEILL, 2010) é que as RNAs são modelos matemáticos que se assemelham às estruturas neurais biológicas e que tem capacidade computacional adquirida por meio de aprendizado e generalização".

Diversas são as aplicações em que as RNAs podem ser utilizadas, dentre elas podemos citar a implementação de controladores não-lineares, implementação de conteúdo endereçável de memória, otimização, classificação de padrões, redução de dimensionalidade, reconhecimento de fala e seleção de informações (MEHRA; WAH, 1992).

Em uma RNA, o aprendizado se dá na fase onde a rede recebe os dados de entrada e ao aplicar os pesos nestes dados, modifica-os a fim de obter conhecimento sobre eles.

O modelo geral do neurônio artificial é composto por suas entradas, que podem ser as saídas para outros neurônios da rede, os pesos, que são responsáveis pelo aprendizado e podem ser ajustados durante o processo, o *net*, que é o somatório de todas as entradas multiplicadas por seus respectivos pesos e a função de ativação do neurônio. Este modelo pode ser visualizado na figura 2.8.

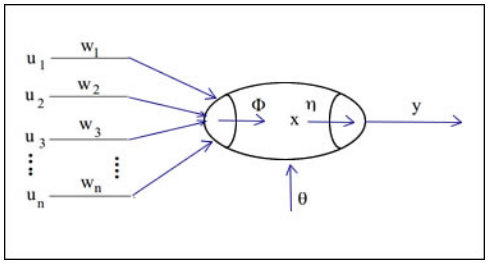

Figura 2.8: Modelo Geral do neurônio artificial. Adaptado de (BARRETO, 2002)

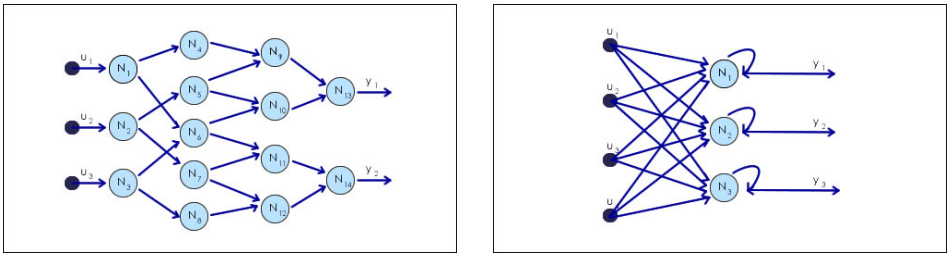

Figura 2.9: Rede neural artificial *feedforward* e rede neural artificial *feedback*. Adaptado de (BARRETO, 2002)

Existem dois tipos de arquiteturas das RNAs, as redes diretas ou *feedforward*, que são caracterizadas por não possuírem ciclos em seu grafo e são comumente representadas em camadas, e as redes recorrentes ou *feedback*, que são redes com realimentação dos neurônios, ou seja, possuem pelo menos um ciclo em seu grafo de conectividade. Para melhor compreensão destes conceitos, na figura 2.9 são apresentadas representações dos dois tipos de redes, a esquerda, uma rede *feed-forward* de quatro camadas em que os dados são introduzidos, processados e conduzidos diretamente à camada saída, sem que haja pontos de retrocesso durante o processo, e a direita uma rede *feedback* de apenas uma camada de processamento em que os dados de saída desta camada são reintroduzidos como entrada na mesma camada para que seja executado um reprocessamento dos mesmos.

2.5 Redes Sociais e Padrões de Dados

Existem muitas formas de se extrair dados a partir de redes sociais, para o *twitter*, rede social abordada nesta pesquisa, são fornecidas oficialmente duas *APIs* de recuperação de dados, a *twitter search API* e a *streaming API*, nas subseções 2.5.1 e 2.5.2 são detalhadas suas principais características e limitações.

Nada impede que as duas *APIs* sejam utilizadas em conjunto, de maneira complementar, na coleta de dados para mineração, uma vez que possuem formas diferentes de busca.

Outras ferramentas de apoio ao monitoramento de redes sociais podem ser encontradas, como o *Google Insights*[5] , para comparação de volumes de buscas por região, *Topsy* [6], para medir a influência de um usuário nas redes sociais baseado no número de menções e compartilhamentos, *Twitter Analyzer* [7], que traça um perfil dos usuários do *twitter*, dentre outras, porém, ambientes como o *Facebook*[8], Orkut[9] e *MySpace*[10] não aceitam *APIs* para a extração de dados (SALUSTIANO, 2010), sendo necessário o desenvolvimento de uma aplicação específica seguindo as normas definidas pelos serviços em questão e autorização de acesso aos dados de cada usuário (RUSSELL, 2011).

2.5.1 *Twitter Search API*

A *twitter search API*, fornece ferramentas para recuperação de postagens recentes de usuários a partir de requisições HTTP por meio do método GET no en-

[5] http://www.google.com/insights/search/
[6] http://topsy.com
[7] http://twitteranalyzer.com
[8] http://www.facebook.com
[9] http://www.orkut.com
[10] http://www.myspace.com

dereço `http://search.twitter.com/search.json?q=` acrescido do parâmetro de busca.

A principal limitação desta *API* é o fato de a busca retornar apenas mensagens recentes, postadas há um período entre 6 e 9 dias anteriores a data da busca. Além disso, ela não permite a realização de buscas complexas, com tamanho maior que 1000 caracteres, não permitindo diversos parâmetros ou parâmetros demasiadamente complexos. Outra limitação é que a *twitter search API* restringe o número de buscas pela complexidade e a frequência das mesmas. Como as solicitações realizadas são anônimas, o limite da taxa é medido pelo IP do cliente solicitante que, ao atingir este limite, passa a receber um erro de HTTP com código 420 informando que a taxa limite de buscas foi atingida. A *twitter search API* não informa exatamente quais são estes limites nem permite que seja feita autenticação ou fornece alternativas para aumento destes limites.

2.5.2 *Twitter Streaming API*

A *Streaming API* fornece ferramentas para buscas atualizadas em tempo real e o acesso ao maior número de mensagens possíveis, porém, contrária à *search API*, é necessária a autenticação com um usuário no *twitter* para ter acesso à seus métodos de busca. Assim como na *twitter search API*, a busca utilizando-se a *API* possui limitações, as principais são a proibição de conexões simultâneas por um mesmo usuário e o bloqueio do usuário e do IP caso o *twitter* considere abusivo o número de tentativas de conexão realizadas em um espaço de tempo.

3 METODOLOGIA

O presente trabalho visa realizar uma pesquisa qualitativa, no que diz respeito a análise de opiniões relacionadas a atitudes e comportamentos dos usuários do *Twitter*, porém com aspectos quantitativos devido aos relatórios de resultados por meio de gráficos de polaridade e métodos estatísticos utilizados nos algoritmos implementados.

A pesquisa apresenta natureza tecnológica, no que diz respeito ao desenvolvimento de um processo para extrair e analisar o sentimento contido em documentos da *web*. Para realização da pesquisa foram utilizadas técnicas de mineração de dados e aprendizado de máquina, bem como o método de aprendizado supervisionado conhecido como SVM (*Support Vector Machine*) que será explicado na subsubseção 2.4.1.

Para extração dos dados do *Twitter*, foram utilizadas *APIs* oficiais disponibilizadas pelo *Twitter*, a *Twitter Streaming API* [1], para captura de mensagens postadas por usuários em tempo real, em conjunto com a *Twitter Search API*[2], para obtenção de mensagens postadas anteriormente ao início do período de coleta.

Visando atender os objetivos propostos, o desenvolvimento do trabalho se dividiu nas etapas descritas abaixo:

- Revisão bibliográfica (set/2011 a dez/2011): Nesta etapa foi realizado um levantamento do estado da arte e analisada a relevância do trabalho no meio científico;

- Estudo dos algoritmos (out/2011 a jan/2012): Inicialmente foi realizado um estudo superficial dos algoritmos de mineração SVM, RNA e os algoritmos

[1] http://dev.twitter.com/docs/streaming-api/

[2] http://dev.twitter.com/docs/using-search

bioinspirados citados no referencial teórico. Em seguida, os estudos foram focados apenas no SVM, que foi o algoritmo utilizado neste trabalho;

- Coleta dos dados (mar/2012);
- Pré-processamento dos dados e mineração (mai/2012 a set/2012): Estas etapas foram realizadas em conjunto com a classificação manual de tweets e os testes, pois durante o desenvolvimento do trabalho, foram necessárias várias alterações na maneira de pré-processar os dados para que se conseguisse uma acurácia satisfatória;
- Análise dos resultados (set/2012): Ao fim da etapa de mineração, foi iniciada a fase de análise dos resultados descritos neste trabalho.

O fluxo das etapas do trabalho pode ser visualizado na figura 3.1:

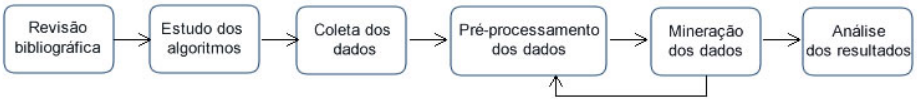

Figura 3.1: Etapas de realização do trabalho

4 ARQUITETURA DO SISTEMA

O sistema utiliza o padrão *MVC* baseando-se na arquitetura cliente-servidor, organizada de maneira em que máquinas clientes executam requisições a um servidor que realiza o processamento destas requisições e retorna as respostas aos clientes. Assim, é possível manter as camadas de modelo e controle no servidor e a camada de visão nos clientes como ilustrado na figura 4.1.

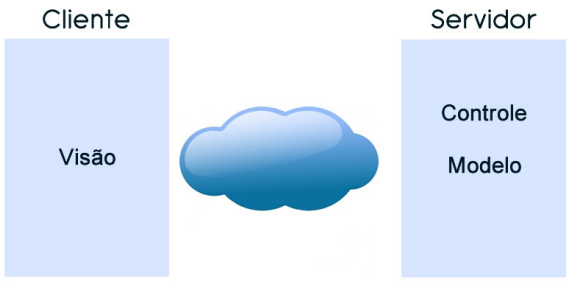

Figura 4.1: Padrão *MVC* baseado na arquitetura cliente-servidor.

(GAMMA *et al.*, 1994) definem *MVC* como uma composição de três tipos de objetos. O modelo é o objeto de aplicação, a visão é a sua apresentação na tela e o controlador é o que define a maneira como a interface de usuário responde às suas entradas. O objetivo deste padrão é desacoplar as camadas que compõem um sistema, de maneira a permitir que sejam criadas diferentes interfaces de usuário para um mesmo modelo sem a necessidade de reescrevê-lo (GAMMA *et al.*, 1994).

Um aspecto importante neste padrão é a definição de um protocolo de comunicação entre a visão e o modelo, permitindo que a visão reflita o status do modelo (GAMMA *et al.*, 1994). O controlador trata os eventos de entrada, encapsulando o mecanismo de resposta à medida que o usuário interage com a interface (visão).

Respeitando essa arquitetura, foram implementados cinco serviços para responder às requisições necessárias ao desenvolvimento do sistema, divididos em

dois serviços para busca de *tweets*, um serviço para realização do pré-processamento dos textos extraídos, um serviço para treinamento do algoritmo de aprendizagem e um último para classificação do conteúdo extraído. A implementação destes serviços está detalhada nas subseções seguintes.

Além destes serviços, foram desenvolvidas classes para representar os *tweets* de maneira simplificada, contendo apenas as informações relevantes para realização da mineração de opinião e classes auxiliares para a execução de tarefas secundárias como a conversão dos objetos retornados pelas *APIs* de busca do *Twitter* e manipulação de arquivos.

Como algoritmo de classificação, foi utilizado o *SVMLearn*, implementação livre e de código aberto desenvolvida por (BODÓ, 2009) do método de aprendizagem de máquina *SVM*.

4.1 Extração dos *tweets*

Para realizar a extração dos dados, o usuário pode optar por duas opções de pesquisa disponíveis no sistema, a busca por *tweets* recentes e a busca em tempo real. Através da busca por tweets recentes, realizada pela *API* de *search* do *Twitter*, é possível recuperar postagens realizadas pelos usuários nos últimos dias. Já a busca por *tweets* em tempo real permite que os textos sejam capturados à medida que são lançados na base de dados do *Twitter* e disponibilizados na *web*.

A figura 4.2 mostra a interface de busca do sistema com as opções disponíveis para execução das pesquisas.

Realizadas as buscas, o sistema armazena os resultados em arquivos de textos com os resultados no formato *JSON (JavaScript Object Notation)*, que apresenta uma estrutura de dados suportada por um grande número de linguagens de programação e de fácil compreensão para humanos (JSON.ORG, 2012).

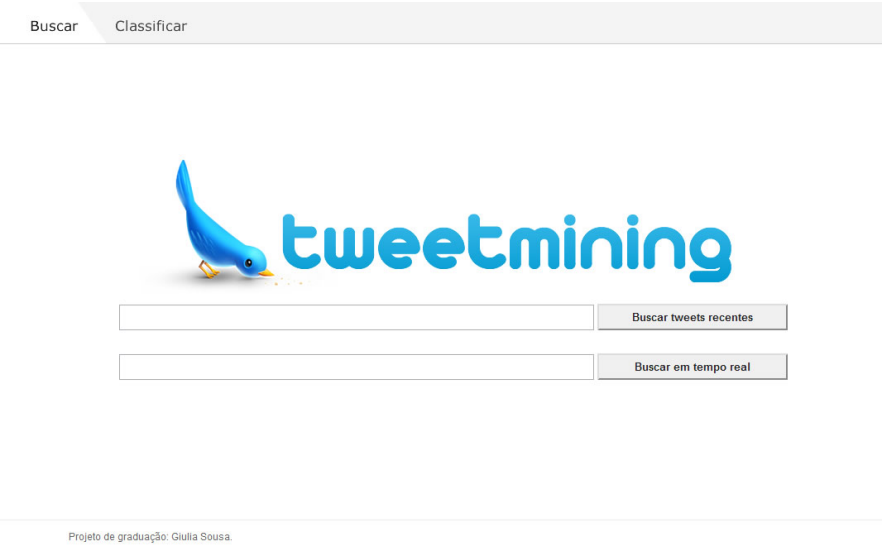

Figura 4.2: Interface de busca do sistema com as opções de busca por *tweets* recentes e busca em tempo real.

4.2 Pré-processamento dos textos

Ao entrar na janela de classificação do sistema, o usuário estará diante das três opções oferecidas, a primeira delas é o pré-processamento dos textos, a partir da qual o sistema recebe um arquivo no formato *JSON* gerado na etapa de extração dos dados e realiza o pré-processamento do mesmo, criando um novo arquivo representando os *tweets* em um vetor de características.

A figura 4.3 apresenta a interface a partir da qual o usuário pode realizar esse processo.

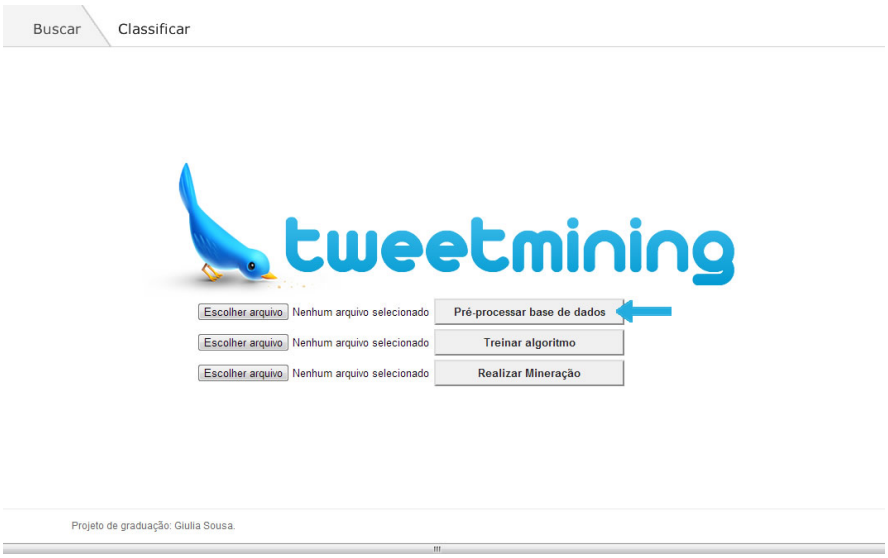

Figura 4.3: Interface para pré-processamento dos *tweets*.

Inicialmente, o *tweet* apresenta uma estrutura complexa, com diversos dados irrelevantes para a mineração efetuada neste trabalho. Considerando isso, os *tweets* foram convertidos para um formato simplificado, contendo apenas os campos significativos para realizar a mineração de opinião proposta.

A seguir apresenta-se um *tweet* em seu formato original:

StatusJSONImpl{ createdAt=MonApr0912: 04: 10BRT2012, id=189368109583437825, text='PontePretaéderrotadapeloXVdePiracicaba-http: //t.co/1D4xhn5b#Campinas', source='web', isTruncated=false, inReplyToStatusId=-1, inReplyToUserId=-1, isFavorited=false, inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=0, wasRetweetedByMe=false, contributors=null, annotations=null, retweetedStatus=null, userMentionEntities=[

], urlEntities=[URLEntityJSONImpl{ start=48, end=68, url=http: //t.co/1D4xhn5b, expandedURL=http: //www.acampinas.com.br/esportes/noticia/ponte-preta-e-derrotada-pelo-xv-de-piracicaba-20120409, displayURL=acampinas.com.br/esportes/notic. . . }], hashtagEntities=[HashtagEntityJSONImpl{ start=69, end=78, text='Campinas'}], user=UserJSONImpl{ id=330553314, name='PortaldeCampinas', screenName='acampinascombr', location='Campinas/SP', description='PortaldeCampinas, jornalismo2.0comnotíciassobrecultura, turismo, gastronomia, educação, esportes, meioambiente, guiadeserviçosemuitomais.', isContributorsEnabled=false, profileImageUrl='http: //a0.twimg.com/profile_images/1523523450/twitter_normal.png', profileImageUrlHttps='https: //si0.twimg.com/profile_images/1523523450/twitter_normal.png', url='http: //acampinas.com.br', isProtected=false, followersCount=280, status=null, profileBackgroundColor='C0DEED', profileTextColor='333333', profileLinkColor='0084B4', profileSidebarFillColor='DDEEF6', profileSidebarBorderColor='C0DEED', profileUseBackgroundImage=true, showAllInlineMedia=false, friendsCount=283, createdAt=WedJul0616: 59: 59BRT2011, favouritesCount=0, utcOffset=-10800, timeZone='Brasilia', profileBackgroundImageUrl='http: //a0.twimg.com/profile_background_images/323039722/bg-twitter-bird.jpg', profileBackgroundImageUrlHttps='https: //si0.twimg.com/profile_background_images/323039722/bg-twitter-bird.jpg', profileBackgroundTiled=false, lang='pt', statusesCount=3865, isGeoEnabled=false, isVerified=false, translator=false, listedCount=3, isFollowRequestSent=false}}

Após a simplificação do tweet, ele passa a ter o seguinte formato:

{"id":189368109583437825,"text":"Ponte Preta é derrotada pelo XV de Piracicaba - http://t.co/1D4xhn5b #Campinas","isoLanguageCode":"pt","createAt":{"year":2012,"month":3, "dayOfMonth":09,"hourOfDay":12,"minute":04,"second":10}}

Essa etapa é realizada para permitir ao algoritmo a obtenção de uma acurácia satisfatória, pois os dados necessitam de um tratamento cuidadoso antes de serem introduzidos às etapas de treinamento e mineração. Para isso, o pré-processamento dos dados foi dividido em quatro etapas descritas a seguir.

4.2.1 Filtragem de *tweets* semelhantes

É normal que os usuários *retweetem* as postagens, ou seja, divulguem a informação em seu perfil na rede social para que seus seguidores tenham acesso a ela, referenciando o autor anterior. Consequentemente, isso gera uma grande quantidade de textos replicados na base coletada. Para resolver esse problema, foi desenvolvido um algoritmo para realizar a filtragem destes textos, identificando os *tweets* idênticos ou muito semelhantes e mantendo uma versão única de cada um. Este algoritmo utiliza as palavras chaves RT e RETWEET para identificar e excluir estes textos.

4.2.2 Dicionário de termos

Um outro problema identificado nos textos extraídos, foi a grande utilização de abreviaturas comumente escritas na *web*. Palavras como “vc”, “tbm”, “ashuashua”, para representar você, também, risos, respectivamente, dificultam o processamento dos dados, pois geram um grande número de termos diferentes com o

mesmo significado. Assim, foi elaborado um dicionário de termos mapeando estas abreviaturas para uma única palavra, com o intuito de melhorar a capacidade de classificação do algoritmo.

4.2.3 Normalização

A normalização tem o objetivo de diminuir o ruído encontrado nos dados, removendo caracteres especiais, palavras sem significado relevante para o algoritmo como preposições, artigos e *hiperlinks*, importantes para a mineração de estruturas na *web*, mas sem significado para este trabalho. Foram utilizadas expressões regulares para identificar e tratar estes ruídos.

A seguir apresenta-se uma versão simplificada do algoritmo utilizado para normalização dos *tweets*:

```
removeInvalidTokens(tweetToProcess){

      processedTweet;

      /*remove accentuation*/
      processedTweet = Normalizer.normalize
         (tweetToProcess, Normalizer.Form.NFKD)
         .replaceAll("\\p{InCombiningDiacriticalMarks}+", "");
      processedTweet = processedTweet.toLowerCase();

      /*remove mentions, hashtags, urls,
      emails and special characters*/
      regexMentions = "^@[^ ]*|[ ]@[^ ]*";
      regexUrls = "^http://[^ ]*|[ ]http://[^ ]*|^ftp://[^ ]*|...";
      regexMail = "^[A-Za-z0-9\\._-]+@ ... ";
      regexSpecialCharacters = "[!-@\\`\\ ... ]";

      regexToRemove;
      regexToRemove.append(regexMentions)
```

```
            .append(regexUrls)
            .append(regexMail)
            .append(regexSpecialCharacters);

        processedTweet.replaceAll(regexToRemove.toString(), "");
        processedTweet.trim();
        processedTweet.replaceAll(" +", " ");

        return processedTweet;

    }

}
```

4.2.4 Conversão para o formato de entrada do *SVM*

Para realizar o treinamento e classificação dos *tweets*, é necessário representá-los de maneira computacional que possa ser entendida pelo algoritmo. Para representá-los dessa forma, foi escolhida a representação por vetor de características, de maneira que os pesos atribuídos a cada termo foi dado a partir do cálculo do *TFIDF (Term Frequency - Inverse Document Frequency)*, uma medida comum de relevância para termos segundo (SALTON, 1989).

O cálculo do *TFIDF* é feito em três passos. Inicialmente, percorre-se todos os documentos construindo um mapa com os termos e sua frequência em todo o *corpus*, ou seja, o número de documentos que contém o termo em questão. Além da frequência no *corpus*, calcula-se a frequência de cada termo dentro do *tweet*, ou seja, a quantidade de vezes que o termo aparece nele, e armazenam-se os termos juntamente com sua frequência indexados ao identificador do *tweet*. Com as frequências calculadas, é necessário calcular a frequência inversa dos termos, que é obtida a partir do cálculo do logaritmo na base 10 do quociente resultante da di-

visão do número de documentos pela frequência do termo no *corpus*. Por último, para conseguir o *TFIDF*, multiplica-se a frequência do termo naquele *tweet* pela frequência inversa do termo.

Com o *TFIDF* calculado, é possível representar o texto no formato *SVMLib*, aceito pelo algoritmo, em que cada *tweet* possui o número +1 ou -1, representando a classe à qual pertence ou 0 quando ela é desconhecida, seguido dos índices de cada termo em ordem crescente e o *TFIDF* calculado para ele. A figura 4.4 apresenta um texto representado nesse formato.

```
+1 2:0.9030899869919435 3:0.0 29:1.7558748556724915
60:3.0909630765957314 61:2.0530784434834195
62:4.5680960720186
```

Figura 4.4: *tweet* representado no formato *SVMLib*.

4.3 Treinamento

Com os dados pré-processados, é possível submetê-los ao algoritmo para realizar o treinamento necessário de acordo com o modelo escolhido. Para isso, o usuário precisa selecionar o arquivo contendo os *tweets* e iniciar o treinamento. Após concluir o treinamento, o sistema mantém o objeto treinado na sessão do usuário, para que, ao executar a classificação de diversas bases não seja necessário um novo treinamento. A interface para execução deste processo é apresentada na figura 4.5.

4.4 Classificação

Após a execução das etapas anteriories, os *tweets* são classificados pelo algoritmo treinado, recuperado da sessão do usuário, e realizadas medidas da acurácia obtida

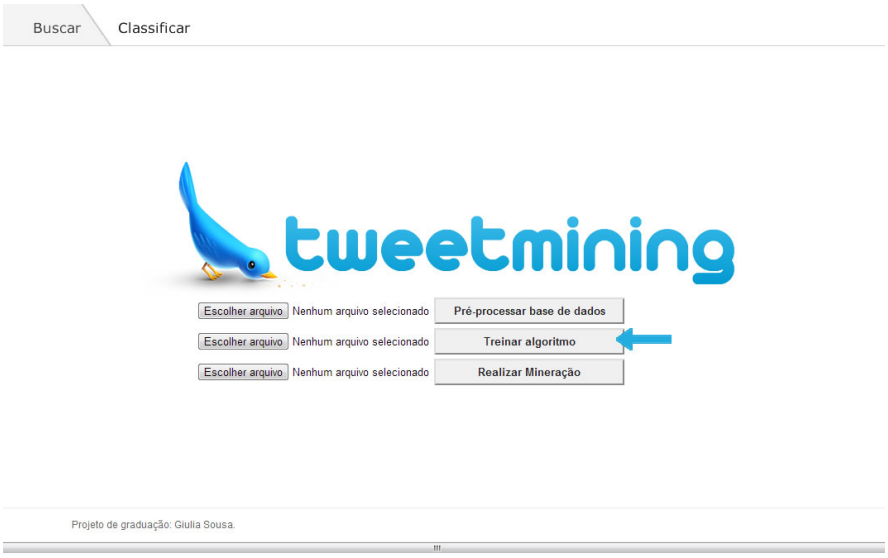

Figura 4.5: Interface para treinamento do algoritmo.

pelo algoritmo a partir dos dados de teste e os resultados da classificação. Para executar essa classificação, o usuário deve inserir o arquivo no formato *SVMLib* já classificado, para que se possa comparar os resultados obtidos pelo algoritmo, pois só assim é possível calcular a porcentagem de acerto durante a classificação. A figura 4.6 apresenta a interface a partir da qual o usuário pode realizar esse processo.

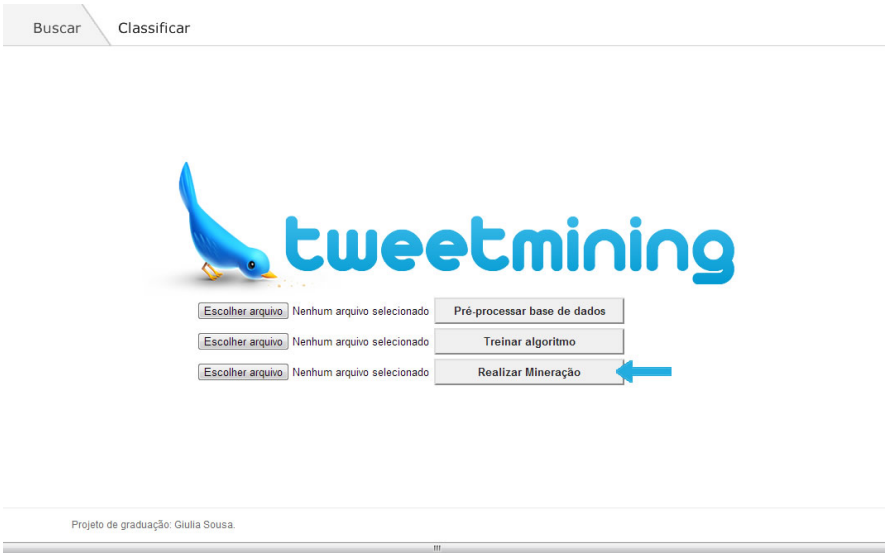

Figura 4.6: Interface para classificação dos *tweets*.

5 RESULTADOS E DISCUSSÃO

O sistema desenvolvido apresentou-se funcional, pois permitiu ao usuário executar todas as etapas propostas atingindo os objetivos de maneira satisfatória. Para obtenção dos resultados, inicialmente foram coletados 110.975 *tweets* falando sobre a cidade de Campinas, dos quais 27% eram textos replicados que foram removidos da base de dados, restando assim 81012 objetos a serem classificados. Destes *tweets*, foram classificados 83% como neutros e os outros 17 % como opinativos. Estes resultados são ilustrados na tabela 5.1.

Número total de *tweets* coletados	110975
Número de *tweets* replicados	29963
Número total de *tweets* classificados pelo algoritmo	81012
Número de *tweets* neutros	67243
Número de *tweets* opinativos	13769

Tabela 5.1: Resultado da primeira classificação realizada pelo algoritmo.

A partir dos 13769 textos classificados como opinativos, foi realizada uma segunda classificação, dividindo os *tweets* em positivos (31%) e negativos (69%). Estes resultados são ilustrados na tabela 5.2.

Número total de *tweets* opinativos classificados pelo algoritmo	13769
Número de *tweets* positivos	4262
Número de *tweets* negativos	9507

Tabela 5.2: Resultado da segunda classificação realizada pelo algoritmo.

Para realização dos testes e validação dos resultados obtidos na classificação, foram selecionados aleatóriamente, uma base contendo 9000 *tweets* opinativos classificados manualmente e uma base de treinamento com 3000 instâncias. Para

realização da coleta de medições, foram realizados dez experimentos utilizando uma mesma máquina de configuração: 4GB de memória RAM e processador Intel Core2Duo com 2 núcleos de processamentos de 3GHz cada e 4MB de cache L2.

As medições coletadas levam em conta as seguintes métricas:

- Porcentagem de acerto quando utilizado na base de testes (%);
- Precisão (π) : mede o quanto a classificação feita pelo algoritmo coincide com a classificação manual feita por um especialista;
- *Recall* (ρ): mede o quanto a classificação feita pelo especialista coincide com a classificação feita pelo algoritmo;
- *F-measure (F)*: representa uma relação de correspondência entre Precisão e Recall;
- Desvio Padrão (σ): mede a variabilidade dos valores à volta da média;
- Tempo médio de treinamento.

As tabelas 5.3 apresenta estas medições.

Tempo médio de treinamento (ms)	86965,3
Acerto (%)	84,2903
Desvio Padrão (σ)	0,06
Precisão (π)	0,7292
***Recall* (ρ)**	0,8165
F-measure (F)	0,7704

Tabela 5.3: Apuração das medições considerando uma base de treinamento de 3000 instâncias.

A partir das medições apuradas, pode-se perceber que de maneira geral, o algoritmo apresentou bons resultados. Apesar do tempo de treinamento se mostrar um pouco alto, pode-se considerá-lo viável devido ao número de instâncias.

No sistema desenvolvido, estes resultados são apresentados de maneira simplificada ao usuário de forma gráfica. A tela de resultados apresenta um gráfico informando a acurácia obtida a partir da base de testes e outro gráfico representando a classificação das instâncias. A figura 5.1 mostra a interface de resultados apresentada ao usuário.

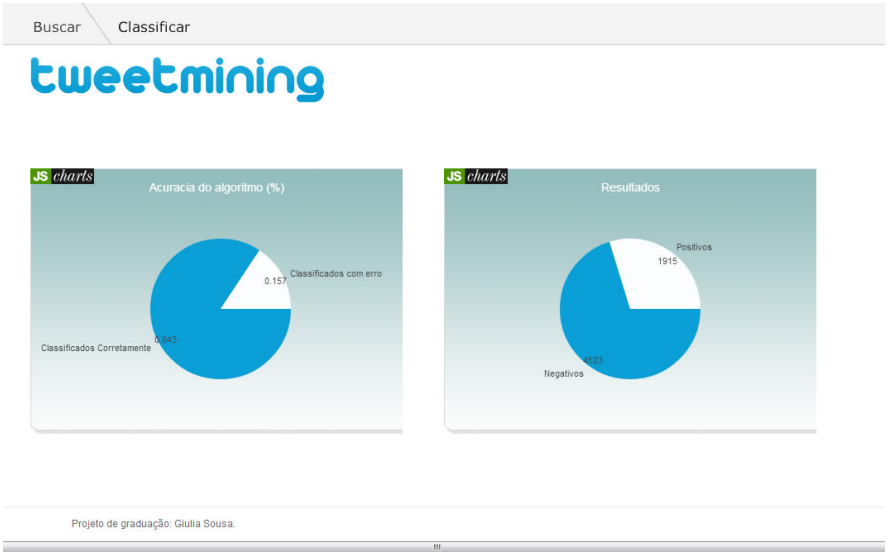

Figura 5.1: Interface de apresentação de resultados da classificação de *tweets* positivos e negativos extraído do sistema.

6 CONCLUSÕES E TRABALHOS FUTUROS

Diversas dificuldades foram encontradas ao longo do desenvolvimento do projeto, pois a interpretação de textos e classificação dos mesmos em meio computacional não é uma tarefa trivial. As principais dificuldades foram problemas referentes à ambiguidades contidas nos textos, excesso de ruídos causados pela má formulação das frases e pelas abreviaturas não oficiais usualmente escritas nos *posts*, base de testes desbalanceada, principalmente ao tratar documentos opinativos e neutros.

Como meio de encontrar melhores formas para solução destes problemas e uma melhor análise dos resultados, foi realizada a classificação manual e estudo de um grande número de documentos. Ao estudar os documentos extraídos sobre a cidade de Campinas, percebeu-se que eles podem ser facilmente separados em outras cinco classes descritas abaixo:

- *Tweets* informando a localização dos usuários: *check ins* de outras aplicações que informam a localização atual do indivíduo;
- *Tweets* sobre política: Como a cidade enfrentava problemas políticos na época de coleta dos textos como troca de pessoas que ocupavam cargos importantes, processos legislativos, entre outros, identificou-se uma grande quantidade de textos expressando comentários referentes à política;
- Ofertas de emprego;
- Eventos e divulgação;
- *Tweets* diversos.

Estes dados são relevantes para uma análise a respeito da cidade, uma vez que contém informações referentes ao nível de satisfação dos habitantes em relação à

seus representantes políticos, o número de ofertas de emprego, que pode ser usado como um indicativo de alerta para uma análise mais aprofundada, além de opiniões e informações referentes aos eventos disponíveis na cidade. A tabela 6.1 apresenta a classificação destes dados:

Classe	**total de *tweets***	**Neutros**	**Opinativos**	**Positivos**	**Negativos**
Localização	4688	4595	93	21	72
Política	6739	944	5795	1562	4233
Ofertas de empregos	879	879	0	0	0
Eventos	1172	890	282	183	99
Outros	15822	15554	268	149	119
Total	29300	22862	6438	1915	4523

Tabela 6.1: Classes de *tweets* identificadas a partir da análise da base extraída.

Analisando os resultados obtidos, pode-se perceber que apesar da grande dificuldade encontrada na classificação de textos extraídos do *Twitter*, devido aos problemas descritos anteriormente e do baixo número de opiniões relevantes a respeito da cidade estudada, em determinados contextos, por exemplo o cenário político de Campinas no período estudado, pode-se obter informações que permitam uma análise interessante e a extração de informações importantes a respeito do cenário.

Os resultados do trabalho foram considerados satisfatórios, uma vez que atingiram os objetivos esperados e encontrou uma solução para o problema proposto. Se aproximaram de resultados encontrados na literatura utilizando métodos semelhantes, como o de (PANG; LEE; VAITHYANATHAN, 2002) que encontrou uma acurácia de 83% para o problema de classificação de polaridade em comentários de filmes e o trabalho de (MATIOLI, 2010) que obteve acurácia de 80.92% atra-

vés do protótipo para mineração de opinião desenvolvido selecionando textos do *Twitter* para testes.

Futuras adições ao trabalho podem envolver a implementação de um método computacional para classificação dos textos de acordo com as cinco classes identificadas [1], visto que elas podem enriquecer de maneira considerável os resultados obtidos.

Outra funcionalidade interessante a ser agregada ao sistema é o aprimoramento da etapa de armazenamento dos dados, uma vez que no sistema implementado os dados são mantidos em arquivos de textos. A utilização de um sistema mais robusto de armazenamento dos dados facilitaria uma análise contínua dos resultados de maneira mais eficiente.

[1] Classes: Localização, política, ofertas de emprego, eventos e outros

REFERÊNCIAS BIBLIOGRÁFICAS

ABASSI, A. Intelligent feature selection for opinion classification. *IEEE Intelligent Systems*, v. 10, p. 72–79, 2010.

ARANHA, C. N.; VELLASCO, M. M. B. R. Uma abordagem de pré-processamento automatico para mineração de textos em português: sob o enfoque da inteligência computacional. Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro. 2007.

BARRETO, J. M. Introdução às redes neurais artificiais. Laboratório de Conexionismo e Ciências Cognitivas UFSC. 2002.

BASTOS, G. M. Algumas aplicações práticas da tecnologia data mining. Sebrae/RJ. 2001.

BATISTA, G. E. A. P. A. Pré-processamento de dados em aprendizado de máquina supervisionado. Instituto de Ciências Matemáticas e de Computação - ICMS - USP. 2003.

BERRY, M. J. A.; LINOFF, G. S. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. [S.l.]: John Wiley & Sons, 2004. ISBN 0471470643.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2003.

BODÓ, Z. *svmlearn. Support Vector Machine Library in Java*. 2009. Disponível em: <http://code.google.com/p/svmlearn/>.

BRABAZON, A.; O'NEILL, M. *Biologically Inspired Algorithms for Financial Modelling*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 3642065732, 9783642065736.

BRACHMAN, R. J.; ANAND, T. Advances in knowledge discovery and data mining. In: *Advances in knowledge discovery and data mining*. [S.l.]: American Association for Artificial Intelligence, 1996. ISBN 0-262-56097-6.

BROWNLEE, J. *Clonal Selection Theory & CLONALG. The clonal selection classification algorithm (CSCA)*. [S.l.], Janeiro 2005.

CATE, F. H. Government data mining: The need for a legal framework. *Harvard Civil Rights-Civil Liberties Law Review (CR-CL)*, v. 43, n. 2, p. 56, Junho 2008.

CHEN, H.; ZIMBRA, D. Ai and opinion mining. *IEEE Intelligent Systems*, v. 10, p. 74–80, 2010.

CIRELO, M. C.; COZMAN, F. G. Aprendizado semi-supervisionado de classificadores bayesianos utilizando testes de independência. Escola Politécnica da Universidade de São Paulo. 2003.

CROFT, B. W.; METZLER, D.; STRHOMAN, T. *Search Engines Information Retrieval in Practice*. [S.l.]: Pearson, 2009.

CURY, G. *Estatísticas de uso da Internet no mundo*. 2010. Disponível em: <http:-//comunicadores.info/2010/02/25/estatisticas-do-uso-da-internet-no-mundo/>.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: *Proceedings of the 12th international conference on World Wide Web*. New York, NY, USA: ACM, 2003. (WWW '03), p. 519–528. ISBN 1-58113-680-3. Disponível em: <http://doi.acm.org/10.1145/775152.775226>.

DIJRRE, J.; GERSTL, P.; R., S. Text mining: Finding nuggets in mountains of textual data. SWSD, IBM Germany. 1999.

ESULI, A.; SEBASTIANI, F. Sentiment quantification. *IEEE Inte*, v. 25, p. 72–75, 2010.

ETZIONI, O. The world wide web: quagmire or gold mine? *Comm*, v. 39, p. 65–68, 1996.

FAYYAD, U. F.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, -, p. 37–54, 1997.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery databases: An overview. *AI Magazine*, v. 13, p. 57–70, 1992.

GAMMA, E.; HELM, R.; JOHNSON, R.; VLISSIDES, J. *Design Patterns. Elements of Reusable Object-Oriented Software.* 1. ed. [S.l.]: Addison Wesley Professional, 1994.

GRAY, P.; WATSON, H. J. The new dss: Data warehouses, olap, mdd, and kdd. Claremont Graduate School, The University of Georgia. 1999.

GROUP, M. M. *Internet World Stats. Usage and Population Statistics*. 2012. Disponível em: <http://www.internetworldstats.com/stats.htm>.

HOTHO, A.; N'RNBERGER, A.; PAAB, G. A brief survey of text mining. University of Kassel. 2005.

IBM. *Coremetrics Intelligent Offer*. 2012. Disponível em: <http://www-142.ibm-.com/software/products/br/pt/personalized-product-recommendations/>.

IBM. *SPSS software*. 2012. Disponível em: <http://www-01.ibm.com/software-/analytics/spss/>.

IMIELINSKI, T.; MANNILA, H. A database perspective on knowledge discovery. *Communications of the ACM*, v. 39, p. 58–64, 1996.

JAIN, R.; PUROHIT, D. G. N. Page ranking algorithms for web mining. *International Journal of Computer Applications*, v. 13, n. 5, p. 0975–8887, Janeiro 2011.

JSON.ORG. 2012. Disponível em: <http://www.json.org/>.

JUNIOR, J. R. C.; PASSOS, E. P. L. Desenvolvimento de uma metodologia para mineração de textos. Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro. 2007.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2st. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2009. ISBN 0130950696.

KANTARDZIC, M. *Data Mining: Concepts, Models, Methods, and Algorithms*. 2nd. ed. [S.l.]: John Wiley & Sons, Inc., 2011. ISBN 0470890452, 9780470890455.

KOSALA, R.; BLOCKEEL, H. Web mining research: a survey. *ACM SIGKDD*, v. 2, p. 1–15, 2000.

LIU, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 3540378812.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma introdução às support vector machines. Instituto de Ciências Matemáticas e de Computação, Universidade Federal de São Paulo. Centro de Matemática, Computação e Cognição, Universidade Federal do ABC. 2007.

MATHEUS, C. J.; CHAN, P. K.; PIATETSKY-SHAPIRO, G. Systems for knowledge discovery in databases. *IEEE Transaction on Knowledge And Data Engineering*, v. 5, p. 903 – 913, 1993.

MATHIAK, B.; ECKSTEIN, S. Five steps to text mining in biomedical literature. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics. 2004.

MATIOLI, L. Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o twitter. Universidade Federal de Lavras. 2010.

MEHRA, P.; WAH, B. W. *Artificial Neural Networks: Concepts and Theory*. [S.l.]: IEE Computer Society Press, 1992.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHEL, T. M. A comparative review of selected methods for learning from examples. In: ____. *Machine Learning: An Artificial Intelligence Approach*. [S.l.]: Tioga Publishing Co, 1983. cap. 3, p. 41–81.

MONTINI, A. d. A. *O poder do data mining*. Setembro 2009. Disponível em: <http://bit.ly/PYd7w9>.

MOREIRA, A.; SANTOS, M. Y.; CARNEIRO, S. Density-based clustering algorithms – dbscan and snn. University of Minho - Portugal. Julho 2005.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc., Hanover, MA, USA, v. 2, p. 1–135, January 2008. ISSN 1554-0669. Disponível em: <http://dl.acm.org/citation.cfm?id=1454711%-.1454712>.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Sentiment classification using machine learning techniques. *Proceedings of EMNLP*, -, p. 79–86, 2002.

PRIOR, A. K. F.; CASTRO, L. N. cpsc: Um algoritmo de enxame construtivo para agrupamento de dados. In: *XVII Congresso Brasileiro de Automática*. [S.l.: s.n.], 2010.

RUSSELL, M. A. *Mining the Social Web*. [S.l.]: O'Reilly Media, 2011.

SALTON, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. [S.l.]: Addison Wesley, 1989.

SALUSTIANO, S. Monitoramento de redes sociais: Muito mais que uma análise de sentimentos. Universidade Castelo Branco. 2010.

SHARMA, P.; TYAGI, D.; BHADANA, P. Weighted page content rank for ordering web search result. *International Journal of Electrical and Computer Engineering Science and Technology*, v. 2, n. 12, p. 7301–7310, 2010.

SIMON, H. A. Why should machines learn? In: MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. (Ed.). *Machine Learning: An Artificial Intelligence Approach*. Berlin, Heidelberg: Springer, 1984. p. 25–37.

SIMONITE, T. Innovation: Software to track our emotional outbursts. *NewScientist Tech*, -, p. –, 2009. Disponível em: <http://www.newscientist.com-/article/dn17101-innovation-software-to-track-our-emotional-outbursts.html>.

SINGH, A. K.; KUMAR, R. A comparative study of page ranking algorithms for information retrieval. *International Journal of Electrical and Computer Engineering*, v. 4, p. 7, 2009.

SRIVASTAVA, J.; DESIKAN, P.; KUMAR, V. Web mining - accomplishments & future directions. In: ____. [S.l.]: University of Minnesota, 2002. cap. 3, p. 51–70.

TERRA. *Suposta censura da Amazon a livros gays gera revolta na web*. Abril 2009. Disponível em: <http://bit.ly/Qu57Dd>.

VAPNIK, V. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982. ISBN 0387907335.

WAGSTAFF, K.; CARDIE, C.; ROGERS, S.; SCHROEDL, S. Constrained k-means clustering with background knowledge. *Proceedings of the Eighteenth Internacional Conference on Machine Learning*, -, p. 577–584, 2001.

WESTON, J.; GAMMERMAN, A.; STITSON, M.; VAPNIK, V.; VOVK, V.; WATKINS, C. *Density Estimation using Support Vector Machines*. [S.l.], 1998.

WILLIAMS, N.; ZANDER, S.; ARMITAGE, G. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *ACM SIGCOMM Computer Communication Review*, v. 36, p. 7–15, 2006.

WITTEN, I. H.; FRANK, E. *Data mining: practical machine learning tools and techniques with Java implementations*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. ISBN 1-55860-552-5.

YI-XING, S.; WEI, W.; ZHEN-HUA, W. Crm measurement indicator system of logistics enterprises for data-mining. In: -. [S.l.: s.n.], 2010. v. 2, p. 1072–1075.