UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC

CURSO DE CIÊNCIA DA COMPUTAÇÃO

LUCAS VALDATI CARDOSO

APLICAÇÃO DA FERRAMENTA SCUP PARA ANÁLISE DE INFORMAÇÃO E MINERAÇÃO DE DADOS EM REDES SOCIAIS (TWITTER)

CRICIÚMA

2014

LUCAS VALDATI CARDOSO

APLICAÇÃO DA FERRAMENTA SCUP PARA ANÁLISE DE INFORMAÇÃO E MINERAÇÃO DE DADOS EM REDES SOCIAIS (TWITTER)

Trabalho de Conclusão de Curso, apresentado para obtenção de Grau de Bacharel do Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientador: Prof. MSc. Paulo João Martins
Coorientador (a): Prof. MSc. Merisandra Côrtes de Mattos Garcia.

CRICIÚMA

2014

LUCAS VALDATI CARDOSO

APLICAÇÃO DA FERRAMENTA SCUP PARA ANÁLISE DE INFORMAÇÃO E MINERAÇÃO DE DADOS EM REDES SOCIAIS (TWITTER)

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Criciúma, 26 de novembro de 2014.

BANCA EXAMINADORA

Prof. Paulo João Martins – Mestre – (UNESC) – Orientador

Profa. Merisandra C. de Mattos Garcia – Mestre – (UNESC) – Coorientadora

Prof. Gustavo Bisognin – Meste – (UNESC)

Prof. Kristian Madeira – Mestre – (UNESC)

RESUMO

As redes sociais estão sendo usadas como fonte de dados de várias pesquisas, seja de negócios, utilidade pública, marketing e nas diversas áreas da computação. Para realização destas pesquisas, as empresas estão utilizando uma técnica chamada mineração de dados, que consiste em uma busca e classificação de grandes quantidades de dados, afim de entender o consumidor e as tendências de mercado. O *Knowledge Discovery in Data Base* ou processo KDD, é utilizado para o processamento destes dados através de padrões como o pré-processamento, que consiste basicamente na limpeza de dados inúteis, a mineração de dados que possibilita a extração e a classificação dos dados através de algoritmos que utilizam técnicas de Inteligência Artificial, e o pós-processamento que é a análise do resultado de toda a extração de dados. No caso deste trabalho, a ferramenta Scup foi a escolhida para realizar todo o processo KDD, desde o pré-processamento até a análise final dos resultados, ela utiliza um algoritmo baseado na técnica de Máquina de Suporte Vetorial, que busca dados a partir de filtros e dimensões. A pesquisa deste trabalho teve o objetivo de analisar as menções relacionadas aos candidatos à presidência do Brasil nas eleições de 2014 no segundo turno. Para isso foram criados os monitoramentos “Aécio Neves” e “Dilma Roussef”, com palavras-chave relacionadas aos dois candidatos. A classificação dos dados ocorreu a partir da análise de cada menção encontrada na base de dados e resultou num resultado praticamente igual aos resultados encontrados nos veículos de pesquisas Datafolha, Ibope e propriamente no resultado final das eleições para presidente do Brasil no segundo turno.

Palavras-chave: Redes Sociais. KDD. Scup. Máquina de Suporte Vetorial. Eleições 2014.

ABSTRACT

Social networks are being used as a data source to several studies, it is for business, public service, and marketing in various areas of computing. To perform this research, companies are using a technique called data mining, consisting of a search and sort large amounts of data in order to understand the consumer and market trends. The Knowledge Discovery in Data Base or KDD process is used to process these data by standards such as pre-processing, which consists essentially in the cleanup of useless data, data mining, which allows the extraction and classification of data through algorithms using artificial intelligence techniques, and post-processing that is the analysis of the results of all data extraction. In the case of this work, Scup tool was chosen to realize the KDD process, from pre-processing to final analysis of the results, it uses an algorithm based on the technique of Support Vector Machine, which fetches data from filters and dimensions. The research of this study was to analyze the citations related to presidential candidates of Brazil in the 2014 elections in the second round. To this were created the monitoring "Aécio Neves" and "Dilma Rousseff" with keywords related to the two candidates. The classification of data was based on the analysis of each term found in the database and resulted in almost the same result to the results found in the Datafolha research vehicles, Ibope and properly in the final result of the elections for president of Brazil in the second round.

Keywords: Social Networks. KDD. Scup. Support Vector Machine. Elections 2014.

AGRADECIMENTOS

Agradeço principalmente a Deus e a toda a minha família que sempre esteve comigo neste período conturbado que é o desenvolvimento do TCC, me apoiando com todas as forças e me ajudando muito. Exalto também a mim mesmo que tive a capacidade de fazer este trabalho praticamente sozinho, mesmo sem nunca ter estudado sobre o tema.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AG	Algoritmos Genéticos
KDD	Descoberta de Conhecimento em Base de Dados
MSV	Máquina de Suporte Vetorial
RNA	Redes Neurais Artificiais

SUMÁRIO

1 INTRODUÇÃO

Redes sociais online têm se tornado extremamente populares, levando ao surgimento e à crescente popularização de uma nova era de aplicações na Web. Associado a esse crescimento, redes sociais estão se tornando um tema central de pesquisas em diversas áreas da Ciência da Computação (BENEVENUTO; ALMEIDA; SILVA, 2011).

A comunicação nas redes sociais surgiu da necessidade que o ser humano tem em compartilhar com o outro, em criar laços que são construídos por afinidades existentes entre eles. Essas afinidades são assuntos em comum, ideais, preferências. Dentre diversos exemplos de assuntos em comum, podemos citar: clubes de futebol, religiões, empresas e músicas. Em meio a tantas formas de interação social, quando essa parte para o ambiente online, é que temos as chamadas redes sociais (ALCANTARA, 2012). Atualmente, as redes sociais são as principais formas de comunicação entre as pessoas no ambiente web, essa necessidade de compartilhar informações entre as pessoas pela Internet desencadeou um aumento significativo de informações.

Devido a esse aumento de informações e a constante presença dos internautas nas redes sociais para expressar suas opiniões sobre produtos, marcas, pessoas, gostos, costumes e assuntos corriqueiros, formou-se um grande interesse por parte de empresas e pessoas em analisar essas informações que podem ser úteis em suas estratégias de Marketing (ALCANTARA, 2012).

A mineração de dados possibilita a busca em grandes bases de dados de informações desconhecidas, permitindo aos gestores uma maior agilidade nas tomadas de decisões. Uma empresa que utiliza Data Mining é capaz de criar parâmetros para entender o comportamento do consumidor, identificando afinidades entre as escolhas de produtos e serviços, prevendo hábitos de compras e analisando comportamentos habituais para detecção de fraudes (PINTO; SANTOS, 2005).

O KDD pode ser visto como o processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração, consistindo na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis, produz uma

relação particular de padrões a partir de dados (FAYYAD; PIATETSKY; SMYTH, 1996, tradução nossa).

De acordo com Fayyad, Piatetsky e Smyth (1996, tradução nossa), existem diversos métodos de Data Mining para encontrar respostas ou extrair conhecimento em repositórios de dados, sendo os mais importantes para o KDD: Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais.

Segundo Han e Kamber (2001, tradução nossa) o conjunto de atividades do processo de descoberta de conhecimento em base de dados é composto de sete etapas: limpeza dos dados, integração dos dados, seleção dos dados, transformação dos dados, mineração dos dados, avaliação dos padrões, apresentação do conhecimento.

A informação e o conhecimento obtidos podem ser utilizados para diversas aplicações, que vão do gerenciamento de negócios, controle de produção e análise de mercado ao projeto de engenharia e exploração científica (HAN; KAMBER, 2001, tradução nossa).

Desta forma, a finalidade desse trabalho constitui-se em analisar informações e minerar dados nas redes sociais utilizando a ferramenta Scup.

1.1 OBJETIVO GERAL

Aplicar a ferramenta Scup para mineração e análise de dados em redes sociais.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos dessa pesquisa baseiam-se em:

a) analisar a ferramenta Scup;
b) utilizar as principais funcionalidades e atributos desta ferramenta, tendo como base a área de Aprendizado de Máquina e o método MSV, que o Scup utiliza;
c) compreender o conceito da Descoberta de Conhecimento em Base de Dados (KDD), bem como seus processos e técnicas;

d) aplicar a mineração de dados utilizando a base de dados da rede social *Twitter*.

1.3 JUSTIFICATIVA

As Redes Sociais estão aproximando milhões de usuários pela Internet. De acordo com a Nielsen (2012b, tradução nossa), as Redes Sociais passaram a frente do *e-mail* como a atividade online mais popular.

Segundo as informações da empresa *Twitter* (TWITTER, 2014, tradução nossa), a rede social tem em sua base mais de 284 milhões de usuários por mês, além disso, são enviados por dia mais de 500 milhões de *tweets* em diversos países, sendo que o *Twitter* exibe suporte a 35 idiomas.

A análise e a extração de conhecimento de redes sociais vêm sendo amplamente utilizadas em várias áreas, incluindo ciências sociais e comportamentais, economia e marketing (WASSERMAN, 1994, tradução nossa).

A aplicação de ferramentas para análise e monitoramento em Redes Sociais em ambientes empresariais possibilita benefícios aos usuários, como por exemplo: A análise dos dados armazenados nos ambientes online onde, a partir desta análise é possível traçar metas e planejamentos de marketing a fim de melhorar a produtividade e a imagem de uma empresa ou do produto comercializado (ALCANTARA,2012).

Através do aumento dessas tecnologias podem ser realizadas tarefas de buscas avançadas, além de relacionamento e interpretação dos dados. A mineração de dados permite que se definam regras de negócio para auxiliar nas tomadas de decisões. Busca-se com isso criar um planejamento das atividades que podem ser desenvolvidas e pensadas a médio e longo prazo, tentando fazer assim uma previsão de tendências futuras baseada no passado (POLITO, 1997).

A grande expansão da área de Mineração de Dados deve-se ao barateamento no armazenamento digital de informações e na expansão do volume de dados armazenados em organizações públicas e privadas (WHESTPHAL; BLAXTON, 1998, tradução nossa).

Dentre os fatores que levaram a escolha da ferramenta Scup está a facilidade de operar os dados resultantes de buscas nas mídias sociais, além de utilizar a tarefa de Aprendizado de Máquina e o método MSV, as configurações do

monitoramento da consulta podem ser avaliadas nas diversas redes existentes, bem como a forma com que desponta os dados, classificando-os em gráficos.

1.4 ESTRUTURA DO TRABALHO

O trabalho é composto por sete capítulos. O primeiro capítulo tem como objetivo mostrar o tema central do projeto e os objetivos selecionados para serem atingidos durante o desenvolvimento, além da definição da justificativa.

A abordagem dos conceitos e regras sobre a descoberta do conhecimento em base de dados (KDD) é o enfoque do capítulo dois, pretende-se com essa abordagem relatar todo o processo para a execução da mineração de dados, que será usado durante o desenvolvimento além de apresentar algumas métodos e técnicas de mineração de dados, tendo como foco principal a área de aprendizado de máquina com o método Máquina de Suporte Vetorial (MSV), que é utilizado neste projeto.

No terceiro capítulo são vistas algumas definições e abordagem das redes sociais, desde a quantidade de usuários ativos em determinadas redes sociais até vantagens e desvantagens da utilização em empresas.

O capítulo quatro é mostrado como é criado um monitoramento para extração de dados em redes sociais, além de apresentar a ferramenta Scup que é a ferramenta de mineração de dados utilizada no trabalho. No capítulo cinco são mostrados alguns exemplos da utilização da mineração de dados utilizando alguma ferramenta.

As etapas do trabalho desenvolvido são listados no capítulo seis. No mesmo capítulo é relatado, por meio de etapas, cada parte desenvolvida. O trabalho realizado também é mostrado, juntamente com os resultados obtidos através dele.

O capítulo sete apresenta a conclusão do projeto e sugestões para trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

O *Knowledge Discovery in Data Base* (KDD) é designado como o processo de extração de conhecimento pela análise de grandes conjuntos de dados, sua principal etapa é o processo de mineração dos dados, que consiste na ação prática de análise e de determinados algoritmos que produzem uma relação de padrões a partir de dados (FAYYAD; PIATETSKY; SMYTH, 1996, tradução nossa). De acordo com Fayyad, Piatetsky e Smyth (1996, tradução nossa), os principais processos de Data Mining para extração de conhecimento em base de dados são: Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais.

Toda a informação e o conhecimento adquiridos podem ser usados para aplicações que vão da condução dos negócios, domínio da produção e diagnóstico de mercado à execução de projetos científicos (HAN; KAMBER, 2001, tradução nossa).

Para desenvolver o processo KDD, segundo Goldschmidt e Passos (2005), devem ser analisados três aspectos: o problema para aplicar ao KDD (Conjunto de Dados, Especialista e Objetivos), os recursos disponibilizados para atender a solicitação (Especialista KDD, Ferramentas e Hardware) e o resultado aplicado ao problema (Modelo do Conhecimento e Histórico do Conhecimento).

Ainda, conforme o mesmo autor, a Descoberta de Conhecimento em base de dados é composta por três etapas: Pré-processamento, Mineração de Dados e Pós-processamento, que são apresentadas na figura 1.

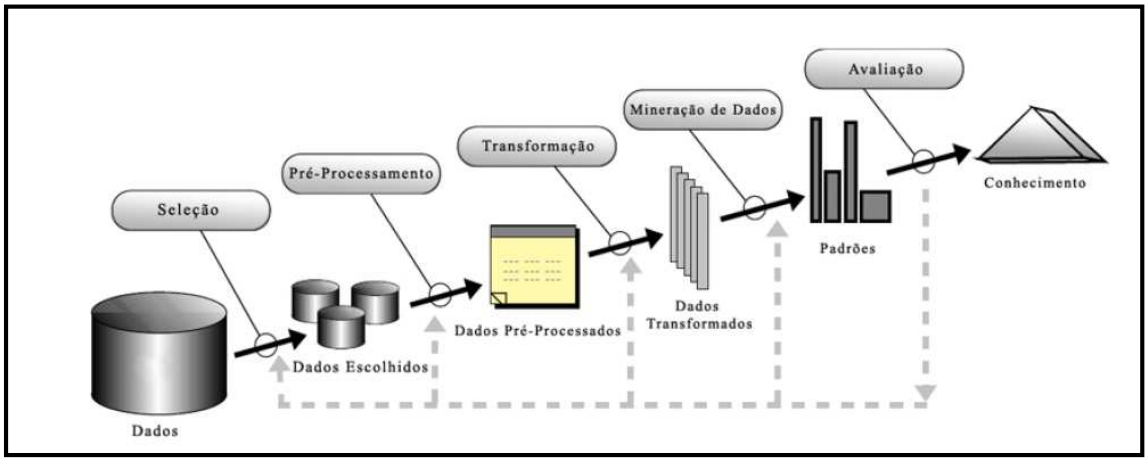
O objetivo do Pré-processamento está relacionado à captação, à organização e ao tratamento dos dados, ou seja, nesta etapa devemos preparar os dados para posteriormente fazer a Mineração dos Dados.

Abaixo, as fases do pré-processamento (GOLDSCHMIDT; PASSOS, 2005):

a) seleção dos dados: nesta fase, deve ser identificado quais as informações de mais valia existem entre as bases de dados que estão à disposição, deve ser escolhido aquelas que irão ser consideradas para o processo de KDD;

b) limpeza de dados: a execução da limpeza dos dados é feita em dados que estejam imperfeitos, ruidosos ou que se contradizem;
c) codificação dos dados: essa etapa escolhe a forma como os dados serão representados no processo KDD. A codificação deve ser específica do algoritmo de Mineração de Dados;
d) enriquecimento dos dados: consiste em acrescentar mais dados aos registros que já existem, apenas para que seja incorporado mais informações para a continuação do processo KDD.

Figura 1 - Processo KDD



Fonte: Fayyad, Piatetsky e Smyth (1996)

Na etapa de Mineração de dados, é feita a aplicação de algoritmos sobre os dados para obtenção do conhecimento, como também são definidas as técnicas e os algoritmos que serão utilizados. Entre as técnicas existentes, podemos citar: Redes Neurais, Algoritmos Genéticos, Modelos Estatísticos e Probabilísticos (GOLDSCHMIDT; PASSOS, 2005).

A escolha de determinada técnica para utilização depende da tarefa de KDD a ser realizada. Podemos destacar algumas tarefas para a Descoberta de conhecimento nesta etapa (GOLDSCHMIDT; PASSOS, 2005):

a) descoberta de associação: está relacionada a procura de padrões e especificações que introduzem a forma como vários itens são agrupados ou acontecem juntos em séries de eventos ou transações;
b) classificação: é feito o mapeamento dos conjuntos de registros em categorias. Encontrada essa função, ela poderá ser inserida aos novos registros, como forma de prever em quais categorias os registros se enquadram;

c) regressão: envolve a pesquisa de uma função que é responsável por mapear os dados que existem no Banco de Dados em valores reais;
d) clusterização: é designado para dividir os registros de uma base de dados em subconjuntos ou clusters. Os elementos de um cluster podem acessar as informações livremente ao cluster que lhe pertence, diferenciando-se dos outros subconjuntos;
e) sumarização: esta ação procura e identifica características comuns que existem entre os conjuntos de dados;
f) detecção de desvios: detecta registros que não competem aos padrões do negócio. São chamados de *outliers*;
g) descoberta de sequência: é uma expansão da tarefa Descoberta de Associação, ela é responsável por pesquisar os itens que são usados repetidamente por diversas transações ocorridas ao longo período.

O pós-processamento compreende o tratamento do conhecimento que foi adquirido na Mineração de Dados, seu objetivo é realizar se existe utilidade no conhecimento descoberto (FAYYAD; PIATETSKY; SMYTH , 1996, tradução nossa).

2.1 TÉCNICAS E TEORIAS DO KDD

Para complementar o processo KDD, deve ser utilizado teorias e técnicas relacionadas ao problema em questão. As seguintes teorias são apresentadas, (GOLDSCHMIDT; PASSOS, 2005):

a) teoria de redes neurais: auxilia o desenvolvimento da Máquina de Suporte Vetorial (MSV);
b) método *rule evolver*: oriundo da Teoria de Algoritmos Genéticos, é aplicado em Classificação e Sumarização;
c) método *NFHB-Class*: combinação das Teorias de Lógica Fuzzy e Redes Neurais, é aplicado em processos de Classificação.

As técnicas e algoritmos são divididos em três relações, segundo Boente (2006):

a) técnicas tradicionais: tecnologias que já existem e que não são relacionadas especificamente com a Mineração de Dados: Redes

Neurais Artificiais (RNA), Lógica Nebulosa, Algoritmos Genéticos (AG) e Estatísticas;

b) técnicas específicas: tecnologias que são especificamente aplicadas no processo de KDD, estas técnicas são fundamentadas por algoritmos computacionais, com o objetivo de buscar uma sequência (passo a passo) para achar a solução do problema;
c) técnicas híbridas: combinam várias técnicas para gerar um sistema híbrido na busca da solução de um problema relacionado a modelagem.

2.2 PRÉ-PROCESSAMENTO

Segundo Goldschmidt e Passos (2005), a fase do pré-processamento está relacionada às funções de captação, organização, ao tratamento e à preparação dos dados para a etapa seguinte (Mineração de Dados). É nesta etapa que serão selecionados os dados que irão ser avaliados em etapas posteriores.

2.2.1 Seleção de Dados

A Seleção de Dados consiste na criação de um conjunto de dados. Nesta etapa, deve ser selecionado um conjunto de dados ou focar em um subconjunto de atributos (variáveis) em que será executada a Descoberta do Conhecimento. Irá variar conforme os objetivos da organização (SASSI, 2006).

Segundo o mesmo autor, na Seleção de Dados é escolhido apenas atributos destacados do conjunto que é analisado na base de dados, ou seja, a Seleção de Dados incide na escolha de um subconjunto de dados relevantes para o objetivo da tarefa. O subconjunto selecionado é aplicado para o algoritmo escolhido para execução da Mineração dos Dados.

2.2.2 Limpeza dos Dados

Nesta etapa é analisado se existe inconsistências nas informações, para corrigir prováveis erros, como também preencher ou eliminar valores desconhecidos ou excessivos, além de eliminar valores que não pertencem ao domínio. O objetivo

geral nessa etapa é a correção dos dados incluídos, eliminando consultas à base de dados dispensáveis, que podem ser executadas posteriormente pelos algoritmos específicos para Mineração de Dados (GOLDSCHMIDT; PASSOS, 2005).

Abaixo estão as principais funções da Limpeza de Dados, segundo o mesmo autor:

a) limpeza de informações ausentes: esta função é responsável por apagar os valores que são ausentes no conjunto de dados. Existem também procedimentos que preenchem valores ausentes: Exclusão de Casos, Preenchimento Manual de Valores, Preenchimento com Valores Globais Constantes, Preenchimento com Medidas Estatísticas e Preenchimento com Métodos de Mineração de Dados;
b) limpeza de inconsistências: é nesta etapa que é identificado e eliminado os valores incoerentes no conjuntos de dados. Existem alguns métodos utilizados para a Limpeza de Inconsistências: Exclusão de Casos e Correção de Erros;
c) limpeza de valores não pertencentes ao domínio: esta função detecta e apaga valores que não fazem parte do domínio de atributos do problema. Alguns métodos utilizados são: Exclusão de Casos e Correção de Erros.

2.2.3 Codificação

O objetivo principal desta etapa é a conversão do conjunto de dados de forma bruta em uma forma padrão de uso. É a fase do pré-processamento responsável pela apresentação dos dados durante o processo de KDD. Neste caso, os dados devem ser codificados de uma forma que satisfaça as necessidades específicas dos algoritmos de Mineração de Dados (GOLDSCHMIDT; PASSOS, 2005).

Existem algumas vantagens que são adquiridas codificando esses dados, são elas: melhor entendimento do conhecimento em questão; redução do tempo de processamento do algoritmo minerador; facilidade do algoritmo tomar decisões globais, já que os valores dos atributos foram englobados em faixas. Todavia, este processo também acarreta desvantagens: diminuição da medida da qualidade de um

conhecimento utilizado, perdendo detalhes que serão relevantes das informações retiradas (SASSI, 2006).

2.2.4 Enriquecimento

É nesta fase que ocorre a associação de valores aos registros que já existem, adicionando mais funções ou termos aos dados disponíveis. A seguir, as principais operações que são utilizadas nesta etapa (GOLDSCHMIDT; PASSOS, 2005):

a) pesquisa: tende a buscar novas informações disponíveis junto às fontes originais;
b) consultas a bases de dados externas: é feita pelo agrupamento de informações fornecidas de outras bases de dados.

2.2.5 Construção de Atributos

O objetivo desta etapa é criar novos atributos com base nos já existentes. Sua importância se dá porque os novos atributos podem apresentar uma semelhança com os atributos que já existem na base de dados, podendo dessa forma, reduzir o fluxo de dados, simplificando o processamento do algoritmo de Mineração de Dados (GOLDSCHMIDT; PASSOS, 2005).

2.2.6 Correção de Prevalência

É utilizada para processos que envolvem a classificação dos dados, ajustando desequilíbrios na distribuição dos registros. A diferença existente entre esta etapa e o processo de Normalização dos Dados é que a Correção de Prevalência não se limita a excluir registros com atributos discrepantes, ela apenas os exclui semanticamente (BATISTA, 2009).

2.2.7 Partição do Conjunto de Dados

É sempre necessário que os dados sejam qualitativos e os resultados obtidos sejam coerentes. Há uma necessidade de os algoritmos que são utilizados

para a geração dessas informações sejam testados. Para isso, deve-se utilizar um conjunto de dados, o qual é um subconjunto do todo que é utilizado para a aplicação de mineração. É necessário particionar o conjunto inicial para submeter ao algoritmo para que ele seja treinado. Além disso, deve-se em seguida, testar novamente o algoritmo com outro subconjunto dos dados, para garantir o resultado.

Os dois subconjuntos escolhidos devem fundamentalmente serem diferentes para que a avaliação seja isenta (BATISTA, 2009).

2.3 MINERAÇÃO DE DADOS

Esta etapa é considerada a principal para o processo de KDD, a Mineração de Dados é responsável pela procura de novas informações, onde os algoritmos são acionados. Cada técnica e algoritmo são provenientes de diversas áreas, como por exemplo: Aprendizado de Máquina, Estatísticas, Redes Neurais, Banco de Dados e diversas outras. A escolha do melhor algoritmo para resolução do problema depende da compatibilidade das aplicações, para obter resultados mais precisos as técnicas mencionadas podem ser combinadas junto dos algoritmos (SASSI, 2006).

2.4 MÉTODOS DE MINERAÇÃO DE DADOS

Segundo Pitoni (1998), a Mineração de Dados é aplicada para analisar um conjunto de dados, tendo a finalidade de descobrir informações que tenham utilidade. São utilizados métodos matemáticos, heurísticos e algoritmos específicos de mineração.

A aplicação dos métodos da Mineração de Dados depende da necessidade da tarefa exposta, que variam de acordo com a extensão da base de dados onde o método será empregado. Existem vários métodos que podem ser utilizados para o pré-processamento dos dados, inclusive é possível fazer uma combinação entre eles, a qualidade do resultado final está diretamente conectada na escolha dos métodos utilizados (GOLDSCHMIDT; PASSOS, 2005).

2.4.1 Redes Neurais Artificiais

Segundo Barreto (2002), a aparição das Redes Neurais Artificiais surgiu pela ideia de tentar imitar a estrutura de um cérebro, através da construção de uma máquina que apresentaria inteligência de acordo com o cérebro original.

As Redes Neurais Artificiais possuem as mesmas características das Redes Neurais Biológicas, elas possuem estruturas neurais compostas por neurônios artificiais.

O neurônio artificial é uma estrutura lógico-matemática que imita as características do neurônio biológico. Na aplicação do neurônio artificial, os dendritos são trocados por entradas, e as relações com o corpo celular artificial são feitas com objetos intitulados como pesos, fazendo a simulação das sinapses. Os estímulos são capturados pelas entradas, logo depois são processados pela função soma, então a função de ativação se encarrega de fazer a substituição do neurônio biológico (BARRETO, 2002).

Conforme o mesmo autor, o desempenho das redes neurais é decorrente de troca de informações entre as unidades de processamento que estão conectadas por canais adjuntos aos pesos.

2.4.2 Métodos Baseados em Redes Neurais

Na Mineração de Dados, cada campo dos registros em uma base de dados envia os dados pré-processados para a entrada do neurônio artificial. A Rede Neural processa as informações lançando uma saída, que corresponde ao objetivo do problema (GOLDSCHMIDT; PASSOS, 2005).

Existem vários exemplos de tarefas de Mineração de Dados que são aptas a serem implementadas por métodos de Redes Neurais, entre elas as principais, segundo Goldschmidt e Passos (2005): Classificação, Regressão, Previsão de Séries Temporais e Clusterização.

Dois algoritmos de aprendizado são destacados abaixo (GOLDSCHMIDT; PASSOS, 2005):

a) *Back-Propagation*: este algoritmo é conhecido como algoritmo de retro propagação do erro, sua principal finalidade é a diminuição do trabalho da função de erro entre a saída gerada pela rede neural e a saída

desejada de acordo com o objetivo. No momento do treinamento, este algoritmo aciona duas etapas, para cada entrada realizada: o processamento para frente e o processamento para trás;

b) *Kohonen*: está relacionado com a classe das Redes Neurais Artificiais Auto-organizáveis. O treinamento desta classe é não-supervisionado, é fundamentado como uma forma de concorrência entre os elementos. As principais aplicações das Redes Auto-organizáveis são: Tarefas de Clusterização e Detecção de Regularidades.

2.4.3 Algoritmos Genéticos

Os Algoritmos Genéticos têm sua base relacionada ao princípio da Seleção Natural de Darwin, para abordagem de vários problemas, em destaque para otimizações. São robustos, simples e prontamente adaptáveis, é uma técnica empregada em diferentes áreas (LUCAS, 2002).

Ainda, segundo o mesmo autor, o Algoritmo Genético segue o objetivo de designar prováveis respostas para o determinado problema a ser solucionado e posteriormente ser submetido ao procedimento da evolução, utilizando as etapas abaixo:

a) avaliação: nesta etapa são avaliadas as soluções, para verificação se realmente elas respondem ao problema referenciado;
b) seleção: são selecionados os indivíduos para reprodução. A chance de uma solução ser escolhida depende da sua capacidade;
c) cruzamento: é feita a recombinação das características das soluções indicadas, fazendo com que novos indivíduos sejam gerados;
d) mutação: são alteradas algumas características dos indivíduos que restaram do processo de reprodução, fazendo com que acrescente variedade à população;
e) atualização: a população recebe os indivíduos que foram criados.
f) finalização: por fim, é verificado se o resultado da finalização da evolução foi atingido. Caso não for satisfatório, volta para a primeira etapa (Avaliação), se foi atingido então finaliza o processo.

Segundo Lucas (2002), a forma de representar os indivíduos é uma etapa crucial no desenvolvimento de um Algoritmo Genético, pois influenciará diretamente na desempenho do programa.

Os termos genoma e cromossoma são utilizados frequentemente como sinônimos de indivíduos em Algoritmos Genéticos. Essa definição indica que um indivíduo seria um amontoado de genes e apresenta um determinado problema: por mais que toda a representação pelo algoritmo seja fundamentada no seu genótipo (conjunto de genes), toda avaliação é feita com o seu fenótipo (conjunto de características do indivíduo que são resultados da decodificação dos genes).

As características mais importantes dos indivíduos são (LUCAS, 2002):

a) genótipo: é a informação que está contida na estrutura de dados dos genes de um indivíduo;
b) fenótipo: é a decodificação do genoma de um indivíduo;
c) grau de adaptação: demonstra a qualidade da resposta para solucionar o problema indicado;
d) grau de aptidão: evidencia a capacidade do indivíduo em analogia à população que o mesmo pertence.

2.4.4 Métodos Baseados em Algoritmos Genéticos

Os Algoritmos Genéticos são aplicados para solucionar problemas complexos que são relacionados à otimização, em resumo, existem vários tipos de Algoritmos Genéticos que são acionados para solucionar problemas complexos, a maior parte deles (Algoritmos Genéticos Simples) realizam a aplicação da evolução de soluções codificadas em cromossomos artificiais (GOLDSCHMIDT; PASSOS, 2005).

Porém, existe ainda alguns outros tipos de Algoritmos Genéticos, desenvolvidos para problemas específicos (MIRANDA, 2011):

a) genitor: o algoritmo é aplicado para preservar os indivíduos de maior aptidão. É um procedimento chamado elitismo, que faz com que a busca pelos elementos seja mais efetiva, gerando maior confiança nos resultados. Basicamente, o funcionamento deste algoritmo é da seguinte forma, os indivíduos selecionados são cruzados com o

parceiro e o resultado é colocado no lugar onde fica o pior indivíduo da população anterior;

b) *Cross Generational elitist selection, heterogeneous recombination ant Cataclysmic mutation* (CHC): o CHC também tem o princípio de selecionar os melhores indivíduos da população atual. Os melhores indivíduos são selecionados tomando como base a população atual e a população criada após o cruzamento entre os indivíduos (cruzamento feito de forma aleatória), fazendo posteriormente a remoção de indivíduos duplos. Assim como no Genitor, os resultados desta busca são extremamente precisos;

c) algoritmos híbridos: estes Algoritmos são utilizados quando os Algoritmos Genéticos não são suficientemente precisos para solucionar problemas de otimização específicos. Os Algoritmos Híbridos utilizam o princípio dos métodos de otimização dos AG's. A busca deste algoritmo é feito por populações, o que muitas vezes pode gerar um tempo maior de processamento.

2.4.5 Métodos Baseados em Instâncias

Também referenciado como aprendizado de casos ou memorização (MOORE, 2005, tradução nossa), neste método os dados já armazenados ficam inalterados e uma função é usada para reconhecimento de quais instâncias dos dados atuais são mais semelhantes às instâncias que serão avaliadas (WITTEN; FRANK; HALL, 2011, tradução nossa).

O funcionamento deste método se dá a partir de uma função que calcula a distância entre a instância que já existe até a nova instância, é verificada a menor distância das instâncias avaliadas e a que tiver o menor valor é transmitida para uma nova instância (MOORE, 2005, tradução nossa).

O *K-Nearest Neighbors* (K-NN) é um dos principais métodos relacionados à Instâncias que utiliza a tarefa de Classificação, o método é simples de entender e implementar e não é necessário treiná-lo para ser aplicado (GOLDSCHMIDT; PASSOS, 2005).

Conforme o mesmo autor, os passos abaixo são executados neste método:

a) primeiramente é feito o cálculo da distância entre o novo registro e os registros já existentes;
b) são identificados os registros com menor distância em relação ao novo registro;
c) apuramento da classe maior frequência entre os registros com a menor distância (resultantes do passo anterior);
d) por fim, é feita uma comparação da classe com maior presença entre os registros com a classe real, identificando erros ou acertos no algoritmo.

2.4.6 Métodos Específicos

Há métodos criados para resolver problemas específicos da Mineração de Dados utilizando determinados algoritmos. O algoritmo mais conhecido, que é voltado para Descoberta de Associações, é o *Apriori* (GOLDSCHMIDT; PASSOS, 2005).

O *Apriori* é um dos algoritmos tradicionais de Mineração de Regras de Associação (AGRAWAL; MIELINSKI; SWAMI, 1993, tradução nossa) . São vários os algoritmos que tiveram sua base inspirada no *Apriori*, são eles: *GPS, DHP, Partition, DIC, Eclat, MaxEclat*, *Clique*, e *MaxClique*, todos eles fundamentam-se no princípio da antimonotonicidade do suporte (GOLDSCHMIDT; PASSOS, 2005).

Ainda segundo o mesmo autor, os algoritmos que foram citados executam basicamente duas etapas:

a) localizam os conjuntos de itens frequentes que atendem à condição de suporte mínimo;
b) quando é localizado o conjuntos dos itens frequentes, são criadas as regras de associação, que devem atender à condição de confiança mínima.

2.4.7 Métodos Baseados em Indução de Árvores de Decisão

Segundo Goldschmidt e Passos (2005), alguns dos principais métodos de Mineração de Dados têm sua referência empregada na criação de árvores de

decisão pelas bases de dados. A ideia de implementação de uma árvore de decisão é a utilização de um modo recursivo de partição de dados.

O sucesso da aplicação das Árvores de Decisão na Mineração de dados pode ser esclarecido nas seguintes menções (TAN; STEINBACH; KUMAR, 2009)

a) fácil interpretação do conhecimento obtido, pois a árvore de decisão utiliza representações gráficas e regras;
b) enfrentamento poderoso contra ruídos;
c) o custo de algoritmos baseados em Indução de Árvores de Decisão não são caros, por mais que exista uma grande base de dados;
d) competência quando trabalha com dados redundante e irrelevantes que podem acarretar diversos problemas no resultado da Classificação.

Abaixo são apresentados os três principais algoritmos para indução de Árvores de Decisão:

ID3 – Foi um dos primeiros algoritmos de indução de Árvores de Decisão. Aplica a recursividade e é um algoritmo fundamentado na busca gulosa, buscando sempre o conjunto de atributos que dividem os exemplos da forma mais perfeita, criando sub-árvores (QUINLAN, 1986, tradução nossa);

C4.5 – o algoritmo é uma melhora do ID3, estão inclusas melhorias como (QUINLAN, 1993, tradução nossa):

a) trabalha com atributos categóricos ou contínuos;
b) dá manutenção em valores desconhecidos, atribuindo a sigla “?” para estes atributos;
c) utiliza um método chamado *gain ratio* para escolher o atributo que tem a melhor resposta do problema;
d) destaca-se em problemas onde há diferença nos custos dos atributos;
e) usa o método pós-poda que faz com que os valores menos expressivos sejam transformados em um nodo folha na árvore.

CART – O algoritmo CART, criado por Breiman et al (1984), utiliza uma técnica não paramétrica que trabalha com árvores de classificação ou árvores de regressão, tendo como base se o atributo é nominal (classificação) ou contínuo (regressão). As principais propriedades deste algoritmo são a facilidade de achar as

relações existentes entre os dados, mesmo quando não são claros, e também a simplicidade e a qualidade como expõe os dados em uma Árvore de Decisão (FONSECA, 1994).

2.4.8 Métodos Baseados em Lógica Nebulosa

A Lógica Nebulosa utiliza um processo que classifica por números uma situação específica. Ela trabalha com variáveis hipotéticas e vagas, tendo o princípio de facilitar a execução e a manipulação de dados dos computadores (SHAW; SIMÕES, 2002).

A Lógica Nebulosa é baseada na teoria dos Conjuntos *Fuzzy*. A teoria dos Conjuntos *Fuzzy* emprega a verificação do grau de pertinência de um elemento, comparando com seu domínio. Esta análise mostra se o elemento verificado realmente pertence ao domínio. Através da aplicação de uma função, o grau de pertinência resulta em um valor real que varia entre 0 à 1, onde 0 é o resultado que o elemento não pertence ao conjunto, e 1 é o resultado que o elemento está contido ao conjunto (KOHAGURA, 2007).

Nos métodos baseados em lógica nebulosa, os registros existentes em bases de dados podem pertencer, ao mesmo tempo, em múltiplos clusters e classes, com variados valores para seus graus de pertinência (GOLDSCHMIDT; PASSOS, 2005).

Um algoritmo conhecido para aplicação de tarefas de Previsão de Série Temporais é o algoritmo de *Wang-Mendel*, é um método utilizado para gerar regras do SIF (Sistema de Inferência *Fuzzy*) de uso genérico e também é aplicado em classificadores *Fuzzy*.

Abaixo estão apresentados os 5 métodos do *Wang-Mendel* (WANG; MENDEL, 1992, tradução nossa):

a) divide o espaço entre a entrada e a saída dos dados numéricos fornecidos pelas regiões *Fuzzy*;
b) gera regras *Fuzzy* a partir dos dados fornecidos;
c) atribui um grau de cada uma das regras providas;
d) cria uma base de regras *Fuzzy* baseada tanto nas regras geradas quanto nas regras linguísticas de especialistas humanos;

e) determina um mapeamento do espaço de entrada para o espaço de saída baseado na base de regras combinadas *Fuzzy*, usando um processo de defuzzificação.

2.4.9 Aprendizado de Máquina

O Aprendizado de Máquina é uma área da Inteligência Computacional que obtém conhecimento por meio de dados (LIBRALAO et al, 2005).

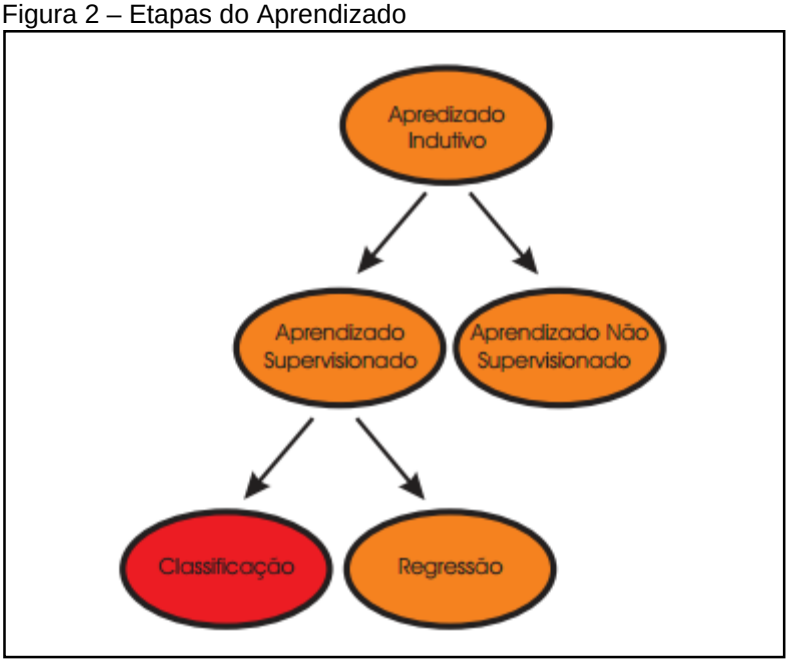
Os métodos de aprendizado de máquina têm o objetivo de alcançar formas de reconhecimento padrão para um conjunto de dados, utilizando a indução (LORENA; CARVALHO, 2007).

O aprendizado de máquina pode ocorrer de duas formas: supervisionado e não supervisionado, conforme pode ser visto na figura 2. Segundo Lorena e Carvalho (2007), no supervisionado há dados de entrada para o treinamento e o objetivo do treinamento é recomendar a melhor saída para as posteriores entradas que ainda não são conhecidas. Já o aprendizado não supervisionado a recepção dos valores é feita apenas para os valores de entrada, não é necessário obter uma saída (WINANDY; FILHO; BENTO, 2007).

Conforme Batista (2003), existe alguns padrões no supervisionado:

a) simbólico: Neste sistema, a aprendizagem se dá a partir da construção de representações simbólicas de um conceito de uma série de exemplos (árvores de decisão, regras de decisão e redes semânticas);
b) estatístico: O foco das técnicas estatísticas é em tarefas onde os atributos usados possuem valores contínuos. A utilização dos modelos estatísticos busca uma aproximação do conceito induzido;
c) *intance-based*: São utilizados casos similares ao que está sendo classificado e é atribuído a classe conhecida como a nova classe;
d) conexionista: Tem relação com a conexão que o cérebro humano faz. Por meio, de uma regressão não-linear, esta ação faz uma aproximação de determinadas funções.

Figura 2 – Etapas do Aprendizado



Fonte: (BATISTA, 2003)

2.4.9.1 Metódo Máquina de Suporte Vetorial (MSV)

A Máquina de Suporte Vetorial é uma técnica de aprendizagem de máquina que cria funções para treinar e classificar os dados (WANG, 2005, tradução nossa). O principal objetivo é obter um equilíbrio entre os erros de treinamento e ao conjunto de teste, minimizando em demasia os ajustes com as amostras de treinamento, aumentando a capacidade de generalização dos dados (VAPNIK, 1999, tradução nossa).

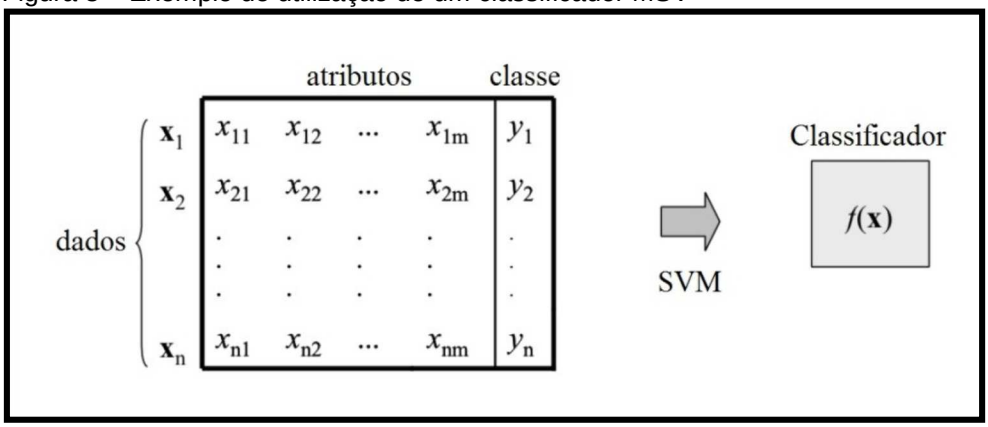
Segundo Campbell e Yiming (2011, tradução nossa), a MSV é uma máquina de aprendizado abstrata, onde é utilizado um conjuntos de dados para treinar e logo após o treinamento, deve ser feito a geração dos dados e fazer uma previsão sem erros do comportamento do restante dos dados.

Este método fundamenta-se nas Teorias de Aprendizagem Estatística e Otimização Matemática. Na Teoria de Aprendizagem Estatística ela procura controlar a geração dos dados, fazendo com que a classificação dos dados seja correta e na Teoria da Otimização Matemática existem funções previamente

definidas que procuram encontrar soluções para o problema (CRISTIANINI; SHAWE-TAYLOR, 2000, tradução nossa).

Conforme Ventura (2010), as MSV empregam a aprendizagem estatística e a indutiva, fazendo com que seja possível buscar uma forma genérica para um determinado grupo de dados, podendo ser exemplificado na figura 3.

Figura 3 – Exemplo de utilização de um classificador MSV



Fonte: Ventura (2007).

2.5 PÓS-PROCESSAMENTO

Na última fase do Processo KDD, o objetivo é analisar e interpretar tudo aquilo que foi gerado pela etapa de Mineração de Dados. Nesta etapa, o especialista em KDD e o conhecedor da aplicação avaliam os resultados que foram originados e determinam novas formas de investigar os dados (GOLDSCHMIDT; PASSOS, 2005).

Existem duas maneiras “gerais” para pós-processar os dados, os métodos subjetivos e os métodos objetivos. No método subjetivo, o usuário deve colocar antecipadamente o conhecimento para o sistema buscar o conjunto de padrões encontrados pelo algoritmo de Data Mining, procurando sempre por respostas que satisfaçam o usuário. Já o método objetivo, não há necessidade de o usuário preencher o conhecimento. A literatura nos diz que o método objetivo é *data-driven* e o subjetivo é *user-driven* (FREITAS, 1999, tradução nossa).

2.5.1 Métodos e Procedimentos Utilizados no Pós-processamento

Existem várias formas de visualizar e contextualizar os dados que foram gerados para os usuários, abaixo estão alguns métodos e procedimentos para executar esta tarefa (Bruha; Famili, 2000, tradução nossa):

a) filtragem do conhecimento: é realizado para aplicações de pós-poda, na utilização de Árvores de Decisão e também em truncagem, nas Regras de Decisão. Na Árvore de Decisão, este método é aplicado quando os algoritmos de indução geram muitas folhas ou regras de decisão exclusivas;
b) interpretação e explanação: este método é utilizado por usuários finais. O conhecimento obtido pode ser exportado para documentos ou alterado para a melhor visualização do usuário;
c) avaliação: os dados podem ser analisados levando em conta vários critérios, como precisão dos dados, inacessibilidade de resultados, complexidade computacional e grau de interesse;
d) integração do conhecimento: combinação e refinamento dos resultados obtidos através de técnicas e outros modelos, para obter maior precisão e aperfeiçoamento dos dados.

2.5.2 Simplificação do Modelo de Conhecimento

Nesta etapa o objetivo é a simplificação da visualização dos dados, fazendo a remoção de alguns detalhes dos modelos de conhecimento, sem perder informações importantes (GOLDSCHMIDT; PASSOS, 2005).

Existem alguns métodos que são utilizados para este caso. Estes métodos têm sua base nas medidas de qualidade das regras, que são a Precisão ou Acurácia da Regra e a Abrangência da Regra (HAN; KAMBER, 2001, tradução nossa):

a) precisão ou acurácia da regra: é um percentual de toda a base de registros que atende a solicitação do antecedente da regra, satisfazendo também o consequente.

b) abrangência da regra: é o percentual dos objetos contidos na base de dados que, no momento de satisfazer o consequente da regra, também satisfaz o antecedente.

3 REDES SOCIAIS

As redes sociais são estruturadas em um conjugado de dois elementos, que são eles: os atores (pessoas conectadas, grupos) que são os “nós” em uma rede, e as suas conexões (ligações e laços de amizade) (WASSERMAN; FAUST,1994, tradução nossa).

Segundo o relatório do primeiro trimestre de 2014 do *Facebook* (FACEBOOK, 2014, tradução nossa), a rede social tem em sua base mais de 802 milhões de usuários ativos em sua rede. A popularidade desta rede social, está diretamente ligada as funções que o usuário pode fazer, desde publicações simples de texto até compartilhamento de vídeos e fotos, entre diversas outras ações. Segundo as estatísticas do *YouTube* (YOUTUBE, 2014), mais de um bilhão de usuários visitam o *Youtube* todos os meses, como também mais de seis bilhões de horas são assistidas por mês, que é o equivalente a uma hora para cada pessoa em todo o mundo. No *Twitter* existem cerca de 284 milhões de usuários ativos por mês e sua principal característica é a postagem de *tweets* de no máximo 140 caracteres de texto puro (TWITTER, 2014).

Segundo a Nielsen (2012a), trinta por cento dos brasileiros acessam pelo menos uma vez ao dia as redes sociais para interagir sobre marcas, produtos e serviços, sendo que 16% deles acessam mais de uma vez ao dia e 14% uma única vez.

As redes sociais online estão sendo utilizadas em grande escala para o marketing viral, fazendo campanhas políticas, divulgando empresas, produtos e diversas outras formas de marketing, onde os usuários compartilham e divulgam o conteúdo com seus amigos virtuais (LESKOVEC; ADAMIC; HUBERMAN, 2007, tradução nossa).

O estudo das redes sociais na Internet tem o foco nas estruturas sociais e como elas surgem, qual é seu tipo, como elas são utilizadas para se comunicar com a rede e como essas trocas de informações são capazes de originar várias formas de conteúdos e dinamismo social (RECUERO, 2009).

3.1 ELEMENTOS NAS REDES SOCIAIS

As redes são relacionadas com metáforas para estudar os sistemas complexos (BARABÁSI, 2003, tradução nossa). Desta maneira, as redes sociais são aplicações das redes para a vinculação com os sistemas sociais, onde os atores são como os nós e as conexões existentes entre os mesmos, como os laços de amizade (DEGENNE; FORSÉ, 1999, tradução nossa). Conforme Recuero (2009), a seguir são destacados os dois elementos existentes nas redes sociais (Atores e Conexões).

3.2 ATORES

São os primeiros elementos em uma rede social. Eles são referenciados como nós (ou nodos). Especificamente, os Atores são as pessoas presentes em uma determinada rede. A função dos atores em um sistema é a interação e a criação de laços sociais entre os outros atores, como forma de moldar as estruturas sociais (RECUERO, 2009).

Ainda, segundo o mesmo autor, os atores são mostrados de um modo um pouco diferente, no início, os atores ficam distantes da interação social, que é o principal meio de comunicação gerenciada pelo computador, ficando difícil de estabelecer quais são os elementos atores. De tal modo, trabalha-se com formas de representar os atores sociais, construindo uma identidade no ciberespaço. Na Internet, o ator pode ser representado por um *blog*, *fotolog*, *twitter* ou mesmo por um perfil no *Facebook*, e mesmo assim, essas ferramentas podem demonstrar um único nó (como um *blog*, por exemplo), que é sustentado por vários atores (um grupo de autores de um *blog* coletivo).

3.3 CONEXÕES

As conexões de uma rede social são as “estradas” que os atores percorrem para trocarem informações. Segundo Recuero (2009), de uma forma geral, as conexões em uma rede social são compostas por laços sociais, desenvolvidos por meio da interação social entre os atores. O principal objeto de estudo das redes sociais são as conexões, pois são elas que alteram as estruturas

dos variados grupos existentes nas redes. A troca de informações na Internet é compreendida devido à possibilidade de conservar os indícios sociais dos indivíduos que trocam informações, navegam na Internet, distribuem e utilizam dados. Todo o histórico das ações utilizadas permanece nas redes. Uma mensagem deixada em um *weblog*, por exemplo, fica estático até que alguém a remova ou o *weblog* deixe de existir. Este mesmo conceito acontece na maioria das interações executadas no computador, onde essas interações são destinadas a permanecerem na Internet, admitindo ao pesquisador a trocar informações sociais mesmo estando apenas ligado remotamente, sem ter o mínimo de conhecimento de onde estas interações estão sendo feitas no tempo e espaço (RECUERO, 2009).

3.3.1 Interação, Relação e Laços Sociais

Segundo Recuero (2009), a Interação, a Relação e os Laços Sociais, são proferidos como elementos de uma Conexão.

A Interação é como se fosse a matéria-prima das Relações e dos Laços Sociais, ela atua como uma rotina sempre comunicacional. A interação é definida, deste modo, como aquela ação que tem uma tarefa de fazer a comunicação entre os elementos e seus pares, fazendo uma espécie de combinação social (WATZLAWICK; BEAVIN; JACKSON, 2000).

Segundo Reid (1991, tradução nossa), a interação social pode ser definida de duas formas:

a) assíncrona: a comunicação entre os elementos nem sempre será em tempo real, a expectativa da resposta não será imediata (*e-mails* e fóruns).
b) síncrona: a expectativa da resposta deverá ser imediata, pois os elementos que estão interagindo encontram-se presentes (*chats on-line* ou troca de mensagens).

O conjunto de interações em um ambiente social, formam as relações sociais (WASSERMAN; FAUST, 1994, tradução nossa). Ainda segundo o mesmo autor, são os padrões das interações que fazem com que as estruturas de uma relação social sejam criadas.

No ambiente da Internet, existem várias formas de relações sociais, pois a troca de informações pode ser ocasionada em diferentes tipos de sistemas e

páginas (troca de informações relacionadas ao trabalho, ambiente pessoal, entre outros). Uma determinada pessoa pode utilizar vários sistemas de interações sociais, como por exemplo, pode utilizar *blogs* para conversação acadêmica, *fotologs* para comunicações mais íntimas e sistemas como o *Facebook* para procurar novas amizades e se relacionar com pessoas conhecidas (RECUERO, 2009).

As relações sociais são relacionadas com a criação dos laços sociais. O laço é uma conexão entre os atores entrelaçados pelas interações, ele implica na solidez das relações formadas pelos objetos. Então, laços sociais são formas de conexão entre os atores, estabelecidos pelo tempo de interações e através da relação social (GARTON; HAYTHORNTHWAITE; WELLMAN, 1997, tradução nossa). O conceito de laço, segundo Wellman (1999, tradução nossa), é que ele é formado por uma ou mais relações, por exemplo, pela proximidade dos indivíduos, ou se existe troca de informações frequentes, pelos fluxos de informação, conflitos ou suportes emocionais. A ligação dos laços implica no recebimento de informações e localizações específicas para a estrutura de um sistema social, os padrões empregados nas relações fazem a organização dos sistemas de troca, controle, dependência, cooperação e conflito.

3.4 TIPOS DE REDES SOCIAIS

A expressão das redes sociais na Internet tem relação com o tipo da ferramenta que é usada pelos atores (os *sites* de redes sociais). Deste modo, as redes sociais consideradas podem ser classificadas de duas maneiras: as redes emergentes e as redes de filiação ou redes de associação (RECUERO, 2009).

3.4.1 Redes Sociais Emergentes

As redes sociais do tipo emergente caracterizam-se pela troca de informações entre os atores sociais. A conexão dos nós entre estas redes se dá a partir das trocas sociais feitas pelas interações sociais e o compartilhamento de informações, tudo isso por meio de um mediador (computador) (RECUERO, 2009).

Segundo a mesma autora, um exemplo de uma rede social emergente é quando analisa-se um *weblog* ou *fotolog* e quando há troca de mensagens feitas

pelos atores das redes, a construção da rede emergente é feita a partir destes comentários, ou seja, na medida que existir interações feitas por um ator e replicado por um outro, a conexão é criada e recriada por meio destas trocas sociais.

Redes emergentes são centralizadas na interação, compostas através da interação mútua entre os atores. A troca de informações de forma mútua gera redes sociais onde os laços são organizados de uma forma pertencente à relação social, que é emergente, caracterizando-os como se “fizessem parte” das relações existentes entre os laços. Além do mais, nesse tipo de rede há concentração de várias trocas de informações entre os mesmos nós (PRIMO, 2003).

3.4.2 Rede de Filiação

As redes de filiação são retratadas como redes de dois modos, por mais que haja apenas um conjunto de atores, isto se dá porque, além da análise do ator, este tipo de rede também leva em conta o conjunto de eventos aos quais os atores especificamente pertencem (RECUERO, 2009).

Ainda segundo a mesma autora, cada evento é um elemento que faz parte da conexão entre o conjunto de atores.

Este tipo de rede é composto de dois tipos de nós: os atores e os grupos, os nós se relacionam por conexões que eles próprios estão contidos (RECUERO, 2009). Segundo Watts (2003, tradução nossa), a rede de filiação permite as pessoas de fazerem interações e troca de informações sem que uma estrutura social esteja montada, pois no momento da criação dos laços sociais a rede de filiação também é criada.

Por fim, as redes sociais de filiação são provenientes das conexões estáticas entre os atores (que são aqueles vínculos conseguidos, por exemplo, por uma lista de pessoas que tem conexão com o “amigo” do ator mas não tem vínculo algum com o ator). Esta ação gera um impacto na rede social, pois novas redes de filiação são criadas no momento da conexão entre os atores (PRIMO, 2003).

3.5 REDES SOCIAIS NAS EMPRESAS

Segundo Rodrigues (2011), com o começo da era social na Internet, as redes sociais como o *Twitter* e o *Facebook*, estão fazendo com que as corporações

e as pessoas estejam cada vez mais presentes em redes sociais. Qualquer forma de comunicação e informação pode trazer vantagens ou desvantagens para as empresas e às pessoas, como uma opinião de um funcionário insatisfeito com a empresa. As redes sociais são um espaço público, qualquer publicação pode ser vista por qualquer pessoa.

3.5.1 Vantagens

Segundo Venancio (2010), abaixo estão inseridas 5 vantagens na utilização das Redes Sociais nas empresas:

a) estar onde a clientela está: as redes sociais contam com diversos clientes das empresas. Desde as pessoas mais novas até as pessoas de mais idade, a utilização das redes sociais se dá diariamente, desde sites de relacionamento como o *Facebook* ou o *Twitter*, até sites de vídeos como o *Youtube*;
b) a empresa fica com uma imagem atual: atualmente as redes sociais são presentes na vida das pessoas. Elas estão associadas a tecnologias novas, a modernidade, ao futuro. Todas estas características podem auxiliar na transmissão da imagem de uma empresa, pois se ela está numa rede social ela deve ser uma empresa dinâmica e atual;
c) interatividade com os clientes: ter uma página da empresa nas redes sociais com possibilidade de receber comentários e opiniões é um canal de interação com os clientes e até com as pessoas que ainda não são clientes;
d) estar nos primeiros lugares nos sites de busca: é necessário que o site da empresa tenha um bom ranking por palavras-chave, com essa vantagem pode atrair mais clientes;
e) ter uma página numa rede social é barato: se for comparado o potencial de visualizações e interações da marca que é gerado pelas redes sociais e com outros meios de divulgação, é possível chegar à conclusão de que as redes sociais são investimentos baratos e de grande valia.

3.5.2 Desvantagens

Além das vantagens, também existem desvantagens na utilização das Redes Sociais nas Empresas (VENANCIO, 2010):

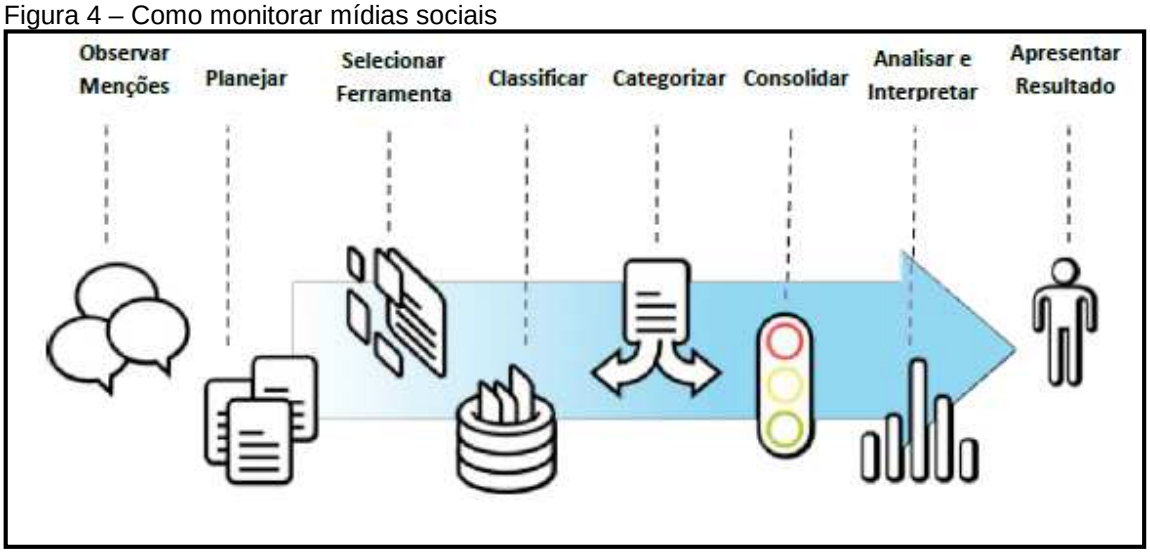
a) estratégia difícil de implantar: existem vários sites de redes sociais, com diferentes focos de público, que utilizam muitas formas de difundir mensagens, tornando-se uma tarefa difícil decidir qual rede social escolher e o que fazer;
b) consome muito tempo: ter uma página em uma rede social exige divulgação e interatividade. Se for criada uma página para a empresa, esta página não pode deixar de se atualizar, e esta tarefa requer tempo;
c) as mensagens divulgadas nas páginas não são totalmente controladas: a maioria dos sites de redes sociais não bloqueiam certos tipos de mensagens e itens publicados. Quando receber um comentário positivo, qualquer pessoa pode ver, o mesmo acontece se um comentário for negativo;
d) lidar com o pessoal e comercial é difícil: a empresa tem que se conscientizar que a maioria dos usuários de redes sociais, tem seu perfil apenas para criação de novas amizades e lazer, a maior parte não tem interesse em ser abordada para uma compra no exato momento. É necessário que as empresas não invadam o espaço dos membros na página.

4 MONITORAMENTO E EXTRAÇÃO DE CONHECIMENTO EM REDES SOCIAIS

O monitoramento de mídias sociais abrange tanto o mercado da comunicação digital quanto as ferramentas e técnicas que são utilizadas por empresas para classificar as informações encontradas. Nos últimos anos houve um surgimento de várias práticas do monitoramento de mídias sociais: empresas especializadas na prática do monitoramento foram criadas, empresas contrataram serviços de monitoramento, lançamento de publicações acerca do assunto e especializações (SILVA, 2012).

Para monitorar mídias sociais, devemos seguir os seguintes passos: coletar previamente os dados, armazenar os dados que foram coletados, classificar, categorizar, adicionar informações e analisar as menções online dos termos definidos, conforme mostra a figura 4 (SILVA, 2012).

Figura 4 – Como monitorar mídias sociais



Fonte: Silva (2012).

Ainda segundo o mesmo autor, o principal objetivo de monitoramento das mídias sociais é a análise da opinião dos consumidores e clientes de um determinado produto ou empresa, dessa forma consegue-se conhecer melhor os públicos alvos, para posteriormente realizar ações que possam satisfazer os clientes e alcançar os objetivos das organizações.

Segundo Salustiano (2010), o monitoramento procura entender o sentimento do consumidor.

O monitoramento faz com que a organização conheça melhor os seus clientes, vigie o mercado de forma profunda, identifica novas tendências, evita denegrir a imagem da empresa e com a análise feita pode conseguir aperfeiçoar e criar novos produtos (MILREU, 2010).

Softwares para monitoramento de mídias sociais são ferramentas que filtram e analisam um conteúdo de texto produzido por mídias sociais. Os conteúdos que as ferramentas encontram são buscados por palavras-chave definidas pelo analista de monitoramento da empresa (LAINE; FRUHWIRTH, 2010, tradução nossa).

Segundo o mesmo autor, são várias as funcionalidades destas ferramentas, elas analisam o volume de dados, verificam o autor da informação, buscam a fonte e por fim exportam as classificações e análises feitas em relatórios e gráficos.

4.1 FERRAMENTA SCUP

Criado em 2009, o Scup é uma ferramenta de monitoramento pleno de mineração de conteúdos em redes sociais, segundo a própria Scup (2014) ela é uma das ferramentas mais utilizadas no Brasil (mais de 22.000 marcas). Ela coleta dados, categoriza e analisa as palavras-chave, que os analistas definem em determinadas mídias sociais. Abaixo apresenta as cadastradas (MAICHAKI; OLIVEIRA, 2012):

a) Google;
b) Twitter;
c) Facebook;
d) Youtube;
e) Flickr;
f) Yahoo;
g) Reclame Aqui;
h) Feed Wordpress;
i) RSS;
j) Slide Share;
k) Linkedin;
l) Vimeo;
m) Tumbrl;

n) Foursquare;
o) Instagram.

A ferramenta monitora em tempo real os dados que foram colocados na pesquisa e após a finalização da mesma, é possível criar relatórios e exportá-los para tabelas e gráficos escolhendo diferentes formas (SCUP, 2014).

Ele utiliza *API's* para a coleta de registros, além de utilizar um algoritmo proveniente do método de SVM para mineração dos dados (MAICHAKI; OLIVEIRA, 2012).

4.1.1 Cadastros, Planos e Preços do Scup

Para empresas, existem quatro planos e preços para implantação, conforme tela abaixo:

Figura 5 – Planos e preços do Scup

Fonte: Scup (2014).

Além de Empresas, também existem planos gratuitos voltados para a área acadêmica, este plano é proposto a professores, orientadores e estudantes de Graduação, Pós-Graduação e Cursos com foco em Marketing, Social Media, Comunicação, Publicidade e Estratégia. Este plano tem o objetivo, de fazer com que se estimule o segmento de mídias sociais, tanto por acadêmicos e professores de instituições de ensino, fazendo com que seja possível compreender sua solução

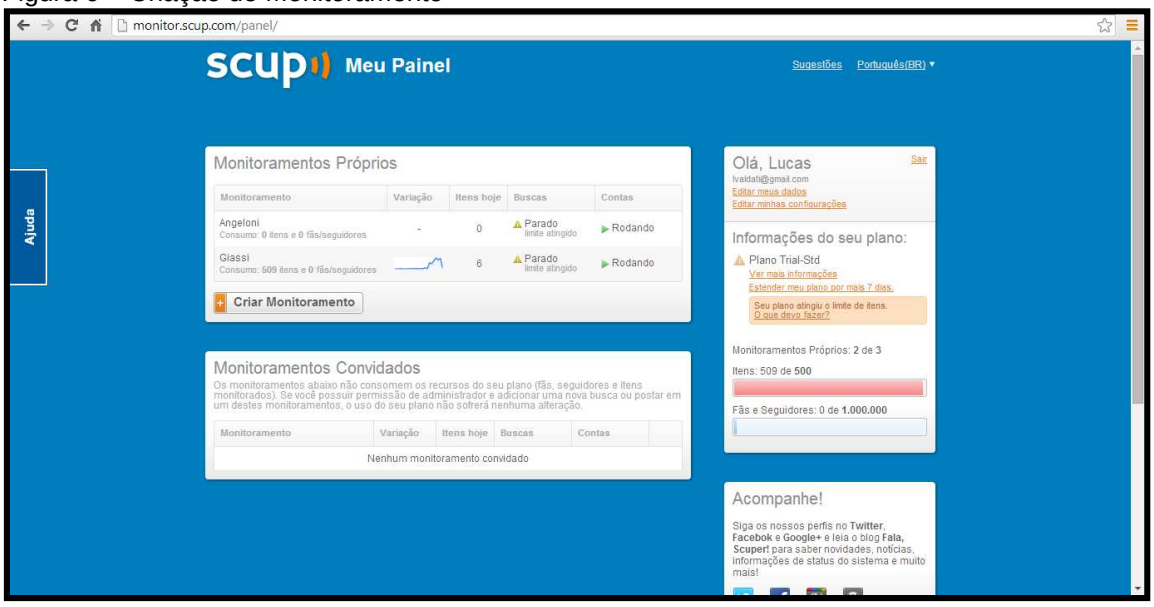
como uma opção de ferramenta de monitoramento nas redes sociais (MAICHAKI; OLIVEIRA, 2012).

Ainda segundo o mesmo autor, em cada plano existe um limite buscas e itens pesquisados. No Plano Acadêmico pode existir até seis monitoramentos e 1000 menções por mês durante um período de dois meses para alunos e seis meses para professores.

4.1.2 Criação de Monitoramento Próprio

Para começar a monitorar os dados depois de criada a conta, na tela de início existe o Cadastro de Monitoramentos, como mostrado na figura 6. É nesta tela que é feito o primeiro passo para a pesquisa, clicando no botão “Criar Monitoramento”.

Figura 6 – Criação de Monitoramento



Fonte: Scup (2014)

É possível criar monitoramentos para Marcas, Concorrentes, Companhias e outros monitoramentos definidos pelo usuário. Logo após o cadastro da Marca que irá ser utilizada na pesquisa, é necessário definir a palavra-chave que será procurada a partir do monitoramento e também em quais mídias sociais será feita a procura dos dados (SCUP, 2014).

4.1.3 Classificação das Menções

Segundo Silva (2012), para fazer a classificação das menções encontradas, deve ser analisado a perspectiva da análise do conteúdo e não simplesmente a mensagem. Para classificar como positivo ou negativo, deve ser levado em conta sempre a perspectiva da marca.

Durante a pesquisa, é possível atribuir sentimentos às menções encontradas, os valores são: “Positivo”, “Negativo” ou “Neutro”, que serão posteriormente mostradas em gráficos e relatórios, conforme Figura.

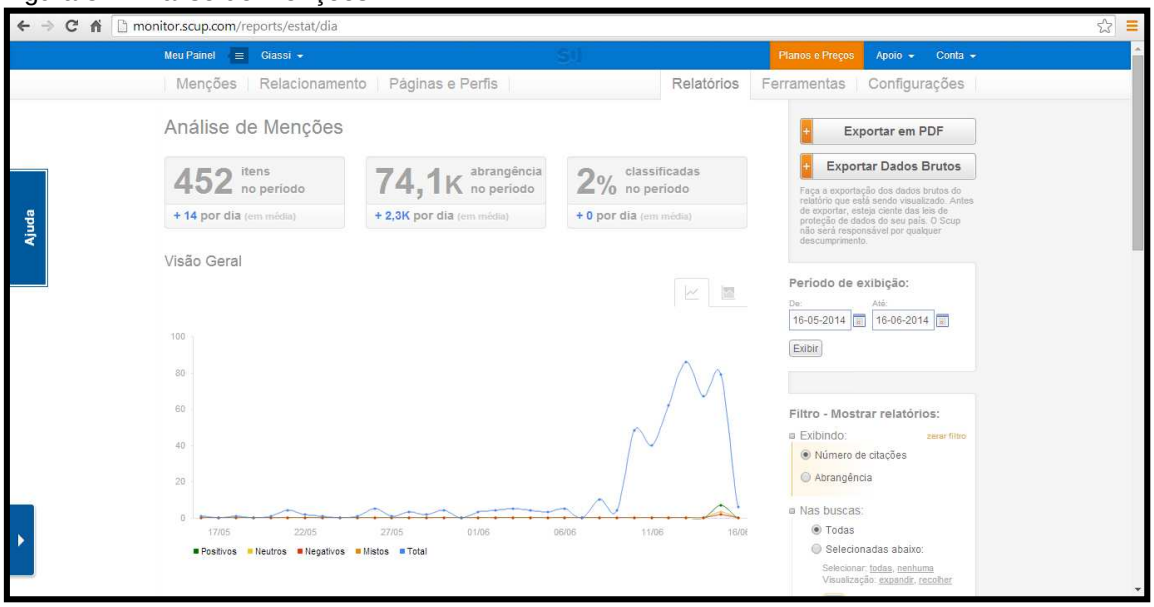
Figura 7 – Atribuição de sentimento

Fonte: Scup (2014).

4.1.4 Geração de Relatórios

Para analisar os dados do monitoramento realizado, o Scup possui algumas formas de geração de relatórios, onde é possível expor os resultados através de menções, termos mais citados, representatividades das redes escolhidas, classificação da marca, comparativos de monitoramentos e outros. Também é possível exportar todos os dados analisados em formato PDF, ou em ferramentas como o Excel (SCUP, 2014).

Figura 8 – Análise de Menções



Fonte: Scup (2014).

5 TRABALHOS CORRELATOS

A mineração de dados em redes sociais pode ser empregada em diferentes áreas para diversos objetivos. A seguir serão descritos alguns trabalhos buscando destacar o foco de algumas das pesquisas.

5.1 MINERAÇÃO DE DADOS EM REDES SOCIAIS

Esta monografia foi desenvolvida por Mariana da Silva Alcantara e apresentada à Faculdade Sete de Setembro – FASETE em 2012. O objetivo é uma análise empírica de três ferramentas de mineração de dados em mídias sociais (Scup, Social Mention e Klout), visando compreender a Descoberta de Conhecimento em Base de Dados (KDD), abordado as fases do Pré-Processamento, Mineração de Dados e Pós-Processamento, e uma abordagem a cerca do tema Redes Sociais.

Nas ferramentas Scup e Social Mention, foi classificado os dados com a palavra-chave “Aborto em Anencéfalos” para posterior análise dos resultados, já na ferramenta Klout, que é uma ferramenta de análise de perfis em mídias sociais, o autor escolheu seu próprio perfil no *Twitter,* para que a ferramenta classifique de 0 a 100, as ações utilizadas nesta mídia social em 90 dias de atividade. Com isso, obtendo o resultado de 14.59.

Por fim, realizou o comparativo da análise empírica das ferramentas, gerando uma tabela com o resultado da usabilidade das ferramentas.

5.2 PROPOSTA DE UMA FERRAMENTA PARA APOIO AO PROCESSO DE KDD BASEADA EM TÉCNICAS DE ASSOCIAÇÃO

Este trabalho foi desenvolvido por Kelven Garcia Campos e apresentado como Trabalho de Conclusão de Curso em Ciência da Computação pela URI – Universidade Regional Integrada do Alto Uruguai e das Missões em 2008. O objetivo foi desenvolver uma ferramenta em Java capaz de abranger as etapas do Processo KDD (Pré-Processamento, Mineração de Dados, Pós-Processamento), pois segundo Campos (2008), existem várias ferramentas que carecem do funcionamento de todas as etapas do Processo KDD, fazendo com que

sejam integradas, a outras ferramentas de apoio para realizar a mineração de dados de forma completa.

5.3 MONITORAMENTO DE MÍDIAS SOCIAIS

Este artigo foi apresentado por Marcos Fernando Maichaki e Robson Leal de Oliveira à PUCPR em 2012 e o objetivo foi fundamentar sobre as mídias sociais e como realizar a mineração de dados e o monitoramento do conteúdo que estão presentes nas redes sociais, bem como um descritivo completo das ferramentas Scup e Social Mention.

Utilizou as ferramentas Scup e Social Mention para analisar as menções dos usuários em palavras-chave já pré-definidas, no caso da Social Mention foi utilizada a palavra-chave “Coca-Cola” e analisou a força da marca, o sentimento de forma geral pela classificação, a paixão que os usuários sentem pela marca (no caso de estarem falando da marca de forma repetitiva), e a medida do alcance dos usuários que estão falando sobre a marca. Já na ferramenta Scup a palavra-chave utilizada foi “Eleições”, que no final da classificação foi exportado um relatório para análise das informações encontradas.

Por fim, os autores definiram vantagens e desvantagens das ferramentas analisadas de acordo com sua utilização durante o trabalho.

6 APLICAÇÃO DA FERRAMENTA SCUP PARA MINERAÇÃO DE DADOS EM REDES SOCIAIS (TWITTER)

Com base no estudo realizado, foi aplicada a mineração de dados em redes sociais utilizando a ferramenta Scup, que se baseia na técnica de Aprendizagem de Máquina e o método MSV para mineração e classificação dos dados, juntamente com todos os conceitos abordados do processo de KDD, desde o pré-processamento ao pós-processamento, para filtrar e encontrar o melhor resultado.

6.1 METODOLOGIA

Com o objetivo de comparar os resultados da mineração de dados feita por meio do monitoramento criado, foi utilizada a base de dados encontrada pela execução da mineração pela ferramenta Scup, utilizando a rede social *Twitter*.

Para o desenvolvimento do trabalho, foram feitas as seguintes etapas:

a) coleta dos dados utilizando os monitoramentos “Aécio Neves” e “Dilma Roussef”.
b) pré-processamento da base de dados extraída do *Twitter*.
c) mineração de dados processados e classificação das menções encontradas com atributo positivo, negativo ou neutro;
d) análise das informações encontradas para comparação por meio de relatórios (pós-processamento).

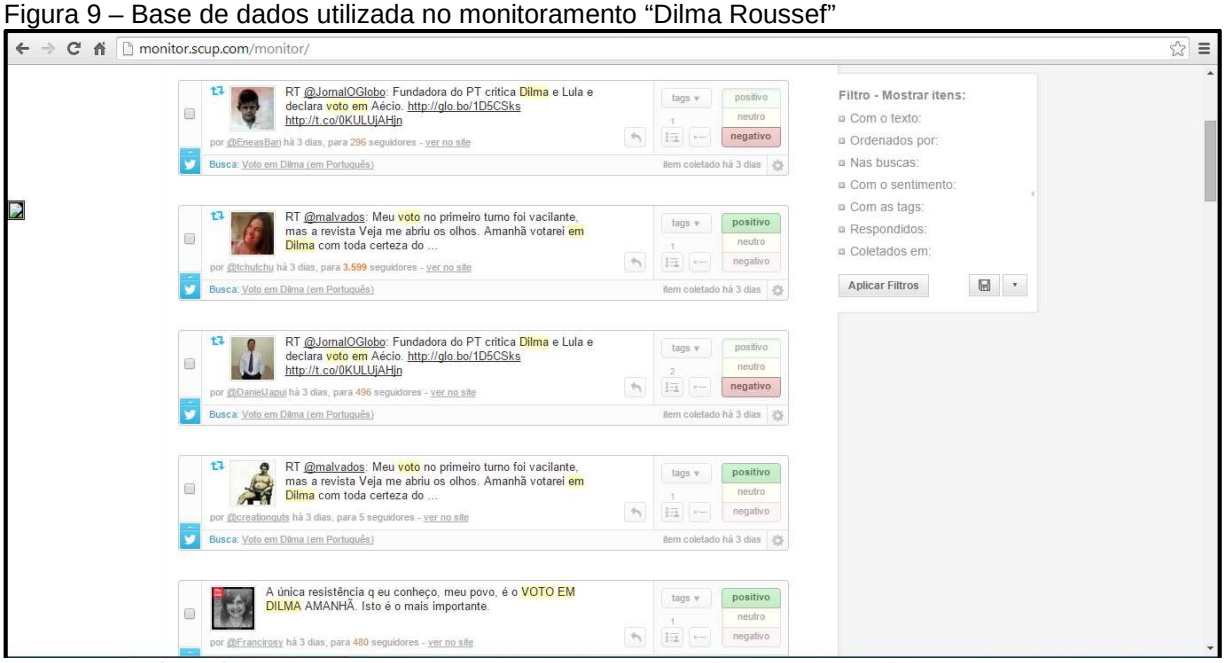
6.1.1 Coleta de Dados

Para a coleta dos dados, criou-se os monitoramentos “Aécio Neves” com as palavras-chave “Aécio Neves” e “Voto em Aécio”, e o outro monitoramento “Dilma Roussef”, relacionando com a palavra-chave “Voto em Dilma”, afim de fazer uma comparação das menções encontradas dos usuários da rede social *Twitter* sobre quem o usuário iria votar no dia 26/10/2014, data da eleição presidencial do Brasil no 2º turno. Por existir muitas menções, os dados foram extraídos apenas no dia 25/10/2014, data que antecedia às eleições.

Conforme já citado anteriormente, a fonte dos dados utilizada foi na rede social *Twitter*, devido à grande quantidade de pessoas que utilizam a ferramenta e também por seus dados seres despontados em textos objetivos, devido à restrição de caracteres imposta pelo próprio *Twitter* (podem ser publicados textos com no máximo 140 caracteres).

A base de dados utilizada pode conter no máximo 2000 registros, restritivo devido à conta do Scup ser para universitários. A figura 9 ilustra as bases de dados coletadas no monitoramento "Dilma Roussef" utilizando a palavra-chave definida (Voto em Dilma).

Figura 9 – Base de dados utilizada no monitoramento "Dilma Roussef"



Fonte: Scup (2014).

Já a figura 10 mostra a coleta de dados feita no monitoramento "Aécio Neves" utilizando as palavras-chave "Voto em Aécio" e "Aécio Neves".

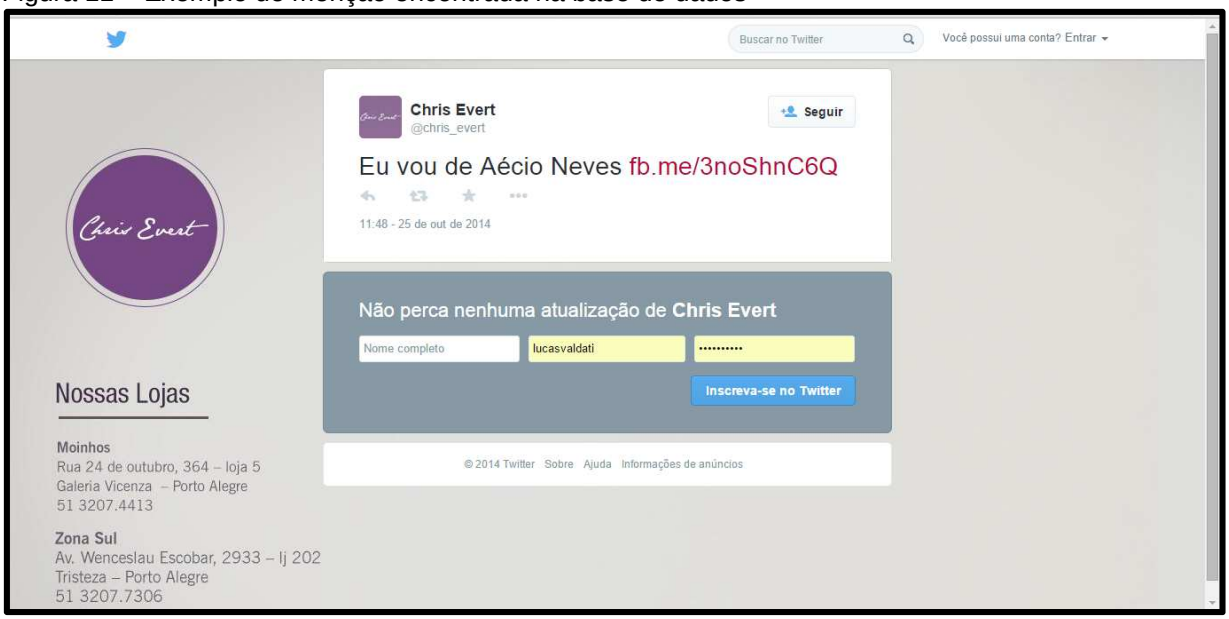
Figura 10 – Base de dados utilizada no monitoramento “Aécio Neves”

Fonte: Scup (2014).

6.1.2 Pré-processamento

Mesmo com o uso de filtros para a extração dos dados monitorados, muitos *tweets* trazem informações irrelevantes para a desenvoltura do trabalho. Um exemplo disso é a postagem: “e a passeata do aécio neves novamente embaixo de minha janela”. Esta postagem foi armazenada devido ao filtro “Aécio Neves” estar sendo utilizado no monitoramento criado, porém este *tweet* não apresenta informação relevante para entrar como parâmetro na comparação dos resultados. No entanto, se verificar o filtro inserido em outro texto, pode ver que ele terá relevância: “Eu vou de Aécio Neves”. É perceptível que no texto o usuário declarou seu voto em Aécio Neves, ou seja, esta informação é relevante para uma futura análise de opinião, conforme pode ser visto na figura 11. Tendo em vista estes argumentos, é indispensável a realização de uma limpeza sobre a base de dados extraída para que não interfira no resultado final.

Figura 11 – Exemplo de menção encontrada na base de dados



Fonte: Scup (2014).

6.1.3 Mineração dos dados

Na etapa de mineração dos dados, ocorre a extração de informação dos dados. Para execução deste processo serão aplicadas técnicas baseadas no algoritmo MSV, disponibilizadas na própria ferramenta, que têm como objetivo à classificação das menções em positivo, negativo e neutro em relação aos candidatos à presidência. Tais dados encontrados por meio dos filtros, fazendo a execução da MSV, ou seja, cada filtro ou palavra utilizada como filtro representa uma dimensão e esta é separada das demais pela utilização do algoritmo empregado.

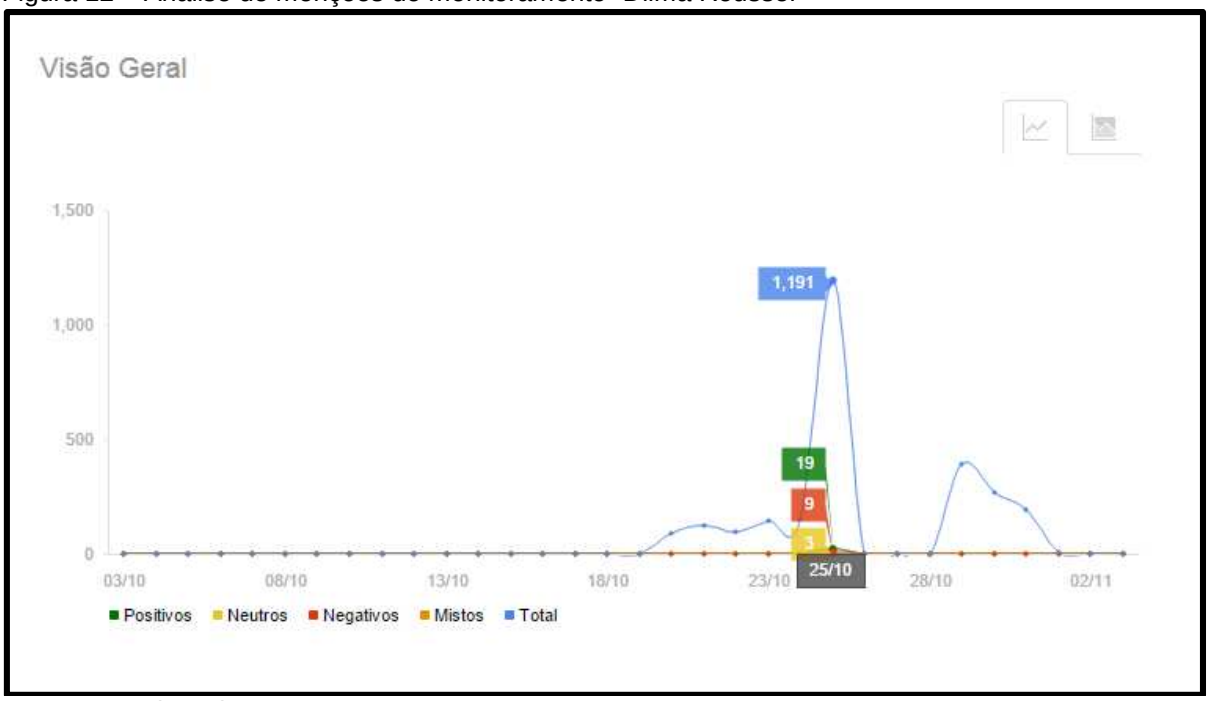
6.1.4 Análise de informação

Por fim, todo o processo executado anteriormente irá ser despontado e comparado por meio de gráficos e tabelas que representam a classificação das menções encontradas na base de dados. Os dados classificados podem ser comparados aos resultados, em alguns veículos de pesquisa do país e propriamente ao resultado oficial das eleições no 2º turno.

O objetivo final é verificar se a pesquisa e a classificação executadas obtiveram uma coerência com o resultado final das eleições.

Um exemplo de um relatório gerado pela ferramenta Scup é o relatório de análise de menções, que tem o objetivo de reunir todos os itens da base de dados do monitoramento e mostrar a quantidade de itens classificados em positivo (verde), negativo (vermelho) e neutro (amarelo). Na figura 12 é possível verificar este relatório gerado para o monitoramento “Dilma Roussef”.

Figura 12 – Análise de menções do monitoramento “Dilma Roussef”

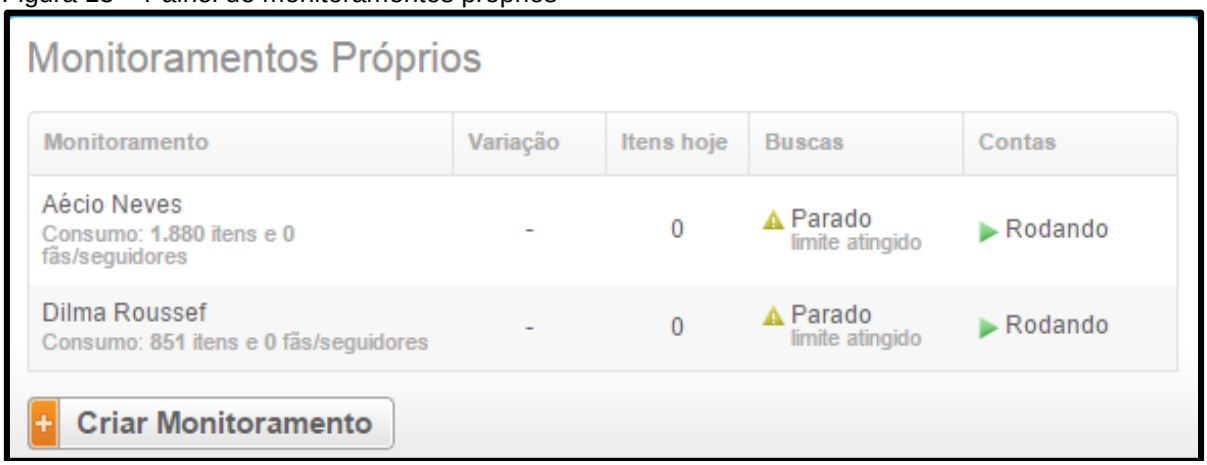


Fonte: Scup (2014).

6.2 RESULTADOS

A base de dados utilizada possuía dois monitoramentos: Aécio Neves (com as palavras-chave Aécio Neves e Voto em Aécio) e Dilma Roussef (com a palavra-chave Voto em Dilma). Após a etapa do pré-processamento que visava a limpeza dos dados que não são úteis como informação, ocorreu a etapa de mineração dos dados ou classificação dos dados, para posteriormente realizar a análise de toda a informação extraída. A figura 13, demonstra os dois monitoramentos criados.

Figura 13 – Painel de monitoramentos próprios



Fonte: Scup (2014)

6.2.1 Coleta

Com o uso da ferramenta de mineração de dados Scup, foram coletados ao todo 2.673 itens no dia 25/10/2014. Deste 1.713 *tweets* do monitoramento “Aécio Neves” e 960 *tweets* do monitoramento “Dilma Roussef”, conforme pode ser visto na figura 14.

Figura 14 – Quantidade de itens coletados

Fonte: Scup (2014).

6.2.2 Pré-processamento

Conforme foi citado em 6.1.2, os dados apresentados possuíam muitas informações irrelevantes. Tendo em vista esta abordagem, estes itens necessitaram de uma análise para não entrar na classificação de dados.

Seguindo a metodologia, a base de dados foi dividida em dois monitoramentos de acordo com o objeto de estudo. Entretanto, nem todos os *tweets* foram classificados, foi retirado uma amostra de apenas 10% do total de itens de cada monitoramento (96 itens do monitoramento Dilma Roussef e 171 itens do monitoramento Aécio Neves).

Para posteriormente executar a mineração e classificação dos dados baseado no algoritmo MSV, a ferramenta utilizada neste trabalho possibilitou a classificação dos itens em até três valores, que foram atribuídas as seguintes regras para cada valor:

a) positivo: menções que claramente votariam no candidato referenciado;
b) neutro: o usuário mostrava alguma informação sobre o candidato, porém não indicava em qual candidato iria votar;
c) negativo: a menção mostrava claramente que a intenção do voto não seria no candidato referenciado.

É possível verificar na figura 15 uma menção pré-processada que ainda não teve sua classificação atribuída em positivo, neutro ou negativo.

Figura 15 – Menção pré-processada

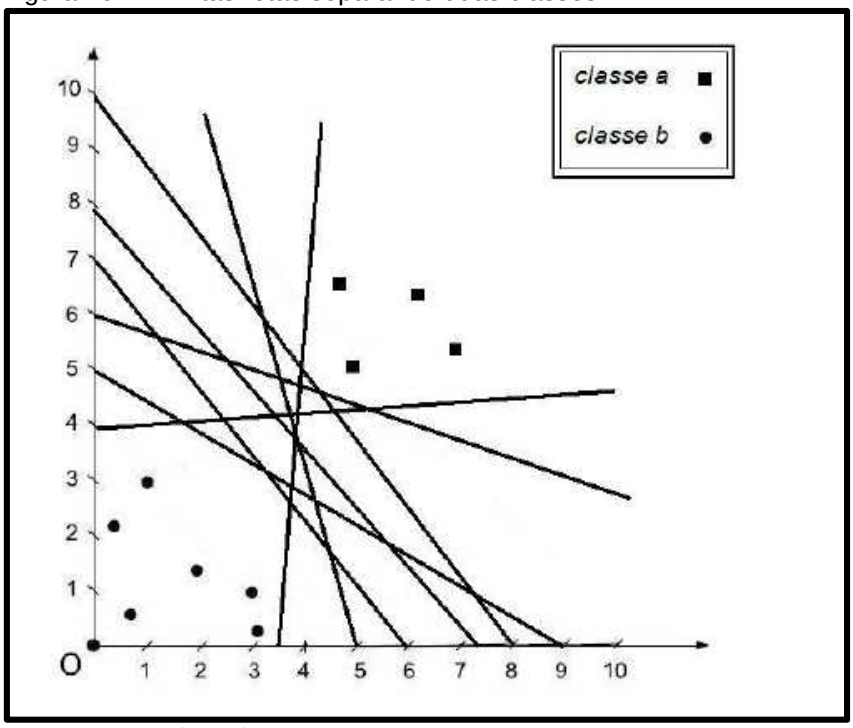
Fonte: Scup (2014)

6.2.3 Mineração

É nesta fase que a classificação dos dados é executada usando o algoritmo MSV. Na ferramenta ele é utilizado na mineração e procura dos dados, fazendo com que cada dimensão (palavra) seja representado por um ponto ou vetor em um espaço *n*-dimensional, traçando um hiperplano que separa a classe das dimensões da melhor maneira possível.

Na figura 16 é possível exemplificar como o MSV trabalha para buscar os dados. No exemplo existem várias retas que separam os dois conjuntos de dados utilizados, ele tentará traçar uma reta que separe os dados da forma mais correta, afim de buscar os dados que mais se aproximem com os filtros utilizados. A reta que é considerada ótima é a que possui a maior capacidade de separar os dados desconhecidos.

Figura 16 – Infinitas retas separando duas classes



Fonte: Santos (2009).

A classificação dos dados também é baseada no algoritmo MSV, como forma de atribuir valores ou sentimentos a cada item extraído, porém na ferramenta, a classificação e atribuição dos valores às menções é realizada manualmente por quem está analisando o monitoramento, por meio da inserção de sentimentos positivos, negativos ou neutros, conforme já comentado.

6.2.3.1 Monitoramento Aécio Neves

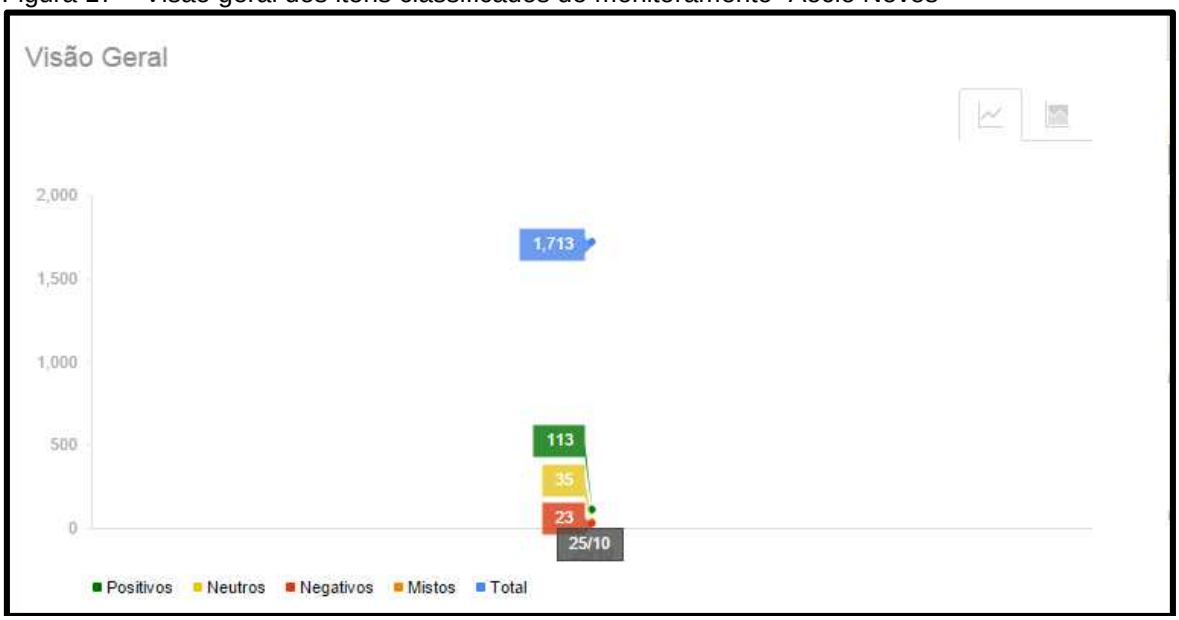
No monitoramento “Aécio Neves” foi adicionado os seguintes filtros para achar os 1.713 itens que compuseram a base de dados:

a) a busca deve ocorrer apenas no dia 25/10/2014, dia que antecede a eleição do 2º turno;

b) apenas a rede social *Twitter* será utilizada como fonte de dados.

Após a restrição da mineração através dos filtros, a figura 17 mostra uma visão geral da quantidade de itens achados durante a pesquisa e dos *tweets* que foram classificados como positivo, negativo e neutro.

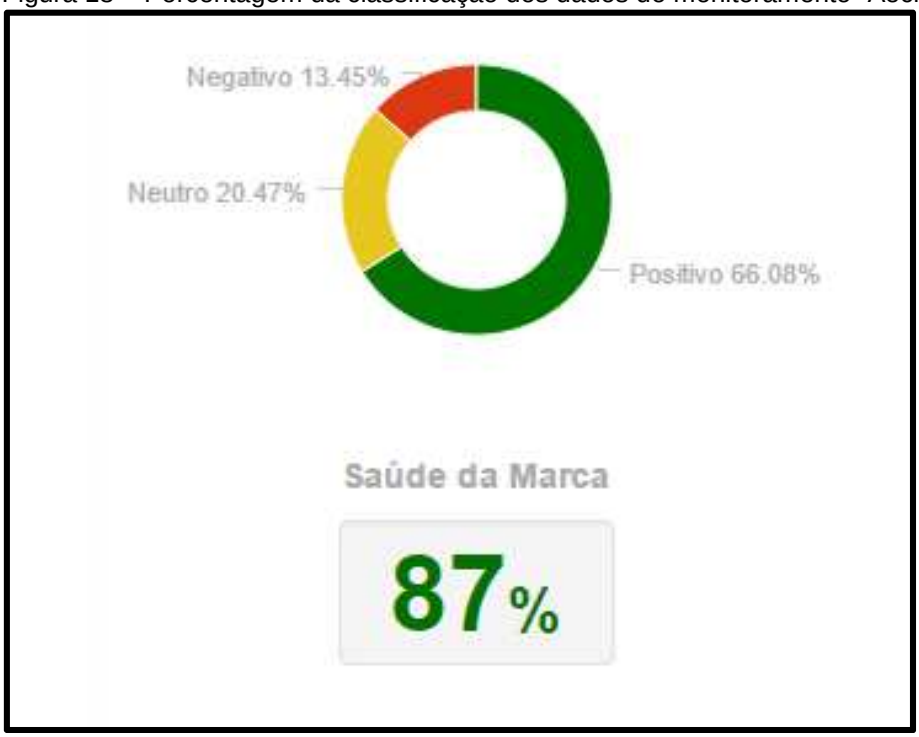
Figura 17 – Visão geral dos itens classificados do monitoramento “Aécio Neves”



Fonte: Scup (2014).

Na figura 18 é mostrado as porcentagens de cada atribuição feita às menções classificadas além de apresentar o campo “Saúde da Marca” que, segundo o Scup (2014), é a porcentagem que representa a proporção de itens que não são negativos. Ela é calculada a partir da soma da quantidade de itens positivos mais a quantidade de itens neutros divididos pela quantidade de itens negativos.

Figura 18 – Porcentagem da classificação dos dados do monitoramento “Aécio Neves”



Fonte: Scup (2014).

Na figura 19 veremos uma nuvem de palavras com os termos mais citados dentro do monitoramento “Aécio Neves”, esta nuvem foi aplicada em todos os itens que estavam contidos no monitoramento.

Figura 19 – Termos mais citados do monitoramento “Aécio Neves”

Fonte: Scup (2014).

De forma geral, 70% dos usuários que mencionaram alguma palavra-chave cadastrada no monitoramento “Aécio Neves” eram do sexo masculino e 30% dos usuários eram do sexo feminino. Além disso, 23,7% dos usuários eram do estado de São Paulo, 14,1% eram do estado do Rio de Janeiro, 7,2% eram do estado de Minas Gerais e 7,1% eram do estado do Rio Grande do Sul, outros estados juntos somaram 47,9%.

6.2.3.2 Monitoramento Dilma Roussef

Assim como no primeiro monitoramento, foram adicionados os seguintes filtros para achar os 960 itens que compuseram a base de dados:

c) a busca deve ocorrer apenas no dia 25/10/2014, dia que antecede a eleição do 2º turno;
d) apenas a rede social *Twitter* será utilizada como fonte de dados.

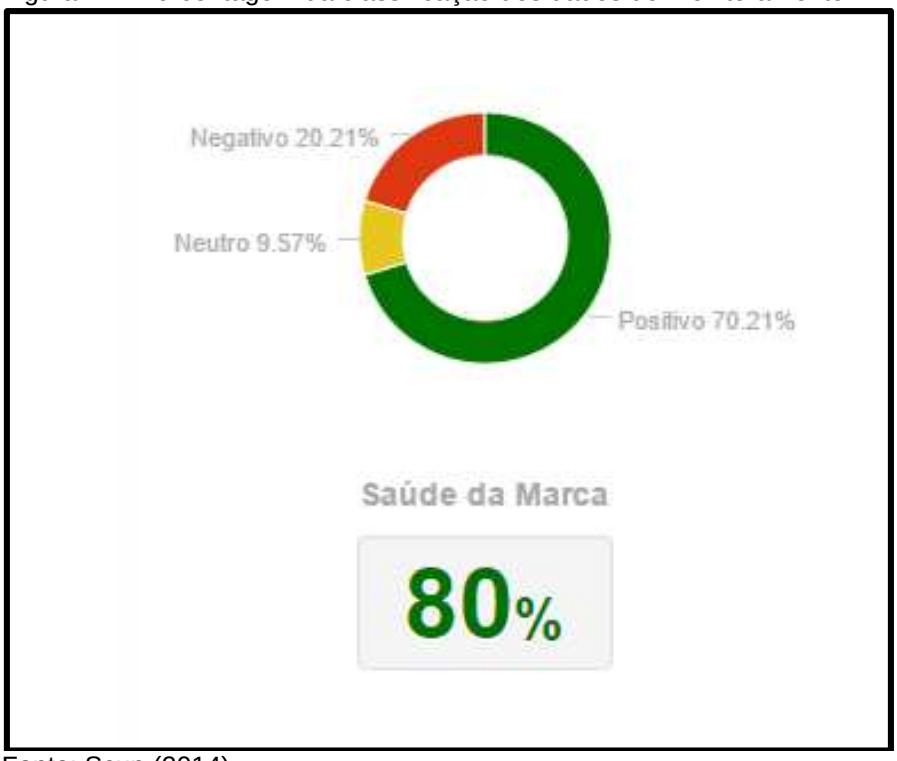
Após a restrição da mineração através dos filtros, a figura 20 mostra uma visão geral da quantidade de itens achados durante a pesquisa e dos *tweets* que foram classificados como positivo, negativo e neutro.

Figura 20 – Visão geral dos itens classificados no monitoramento “Dilma Roussef”

Fonte: Scup (2014).

Na figura 21 é mostrado as porcentagens de cada atribuição feita às menções classificadas além de apresentar o campo “Saúde da Marca”, conforme visto também no monitoramento “Aécio Neves”.

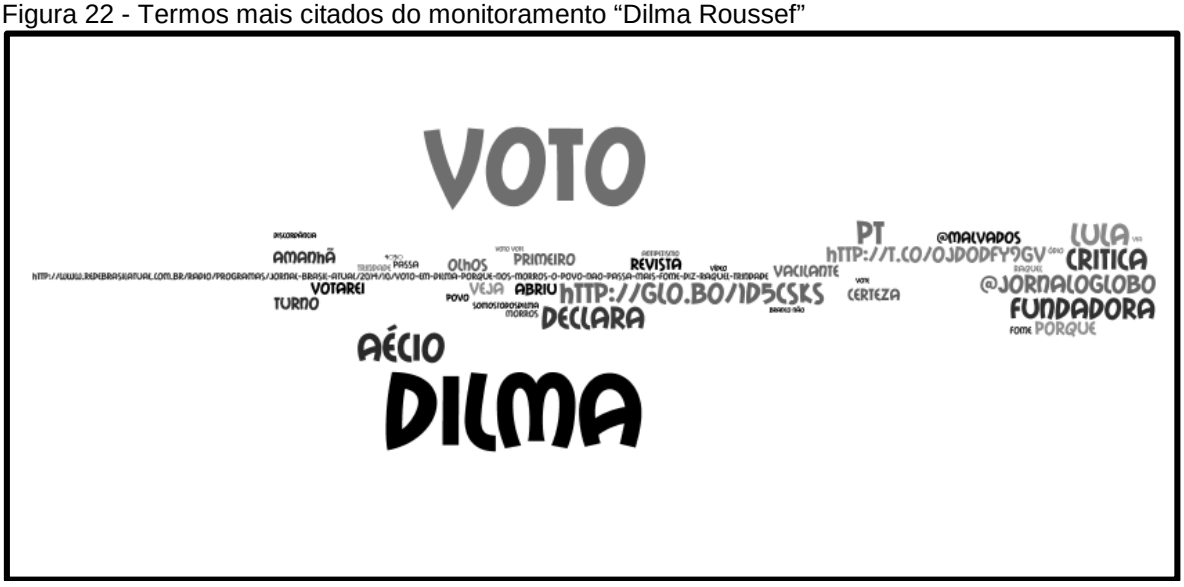
Figura 21 - Porcentagem da classificação dos dados do monitoramento “Dilma Roussef”



Fonte: Scup (2014).

Na figura 22 veremos a nuvem de palavras com os termos mais citados dentro do monitoramento “Dilma Roussef”, assim como no primeiro monitoramento, esta nuvem foi aplicada em todos os itens.

Figura 22 - Termos mais citados do monitoramento “Dilma Roussef”



Analisando todos os itens, 66% dos usuários que mencionaram alguma palavra-chave cadastrada no monitoramento “Dilma Roussef” eram do sexo masculino e 34% dos usuários eram do sexo feminino. Além disso, 21,7% dos usuários eram do estado de São Paulo, 14,3% eram do estado do Rio de Janeiro, 9,4% eram do estado do Rio Grande do Sul e 9,2% eram do estado de Minas Gerais, outros estados juntos somaram 47,9%.

6.2.4 Análise de informação

Com o término da etapa de classificação e mineração dos dados, foi realizada a etapa de comparação com veículos de pesquisa de âmbito nacional (Datafolha e Ibope) e comparação entre os próprios monitoramentos criados, além de comparar com o resultado oficial da eleição presidencial do Brasil.

Com relação aos resultados obtidos, é possível perceber que ambos os monitoramentos tiveram seus resultados parecidos, foram classificados 10% do total de itens de cada monitoramento e, contando apenas os itens com atribuição positiva, a candidata Dilma Roussef apresentou um percentual de aproximadamente

4% a mais do que o candidato Aécio Neves (70,21% a 66,08%). A figura 23 mostra a quantidade dos itens positivos.

Figura 23 – Quantidade de itens positivos na comparação dos monitoramentos

Fonte: Scup (2014).

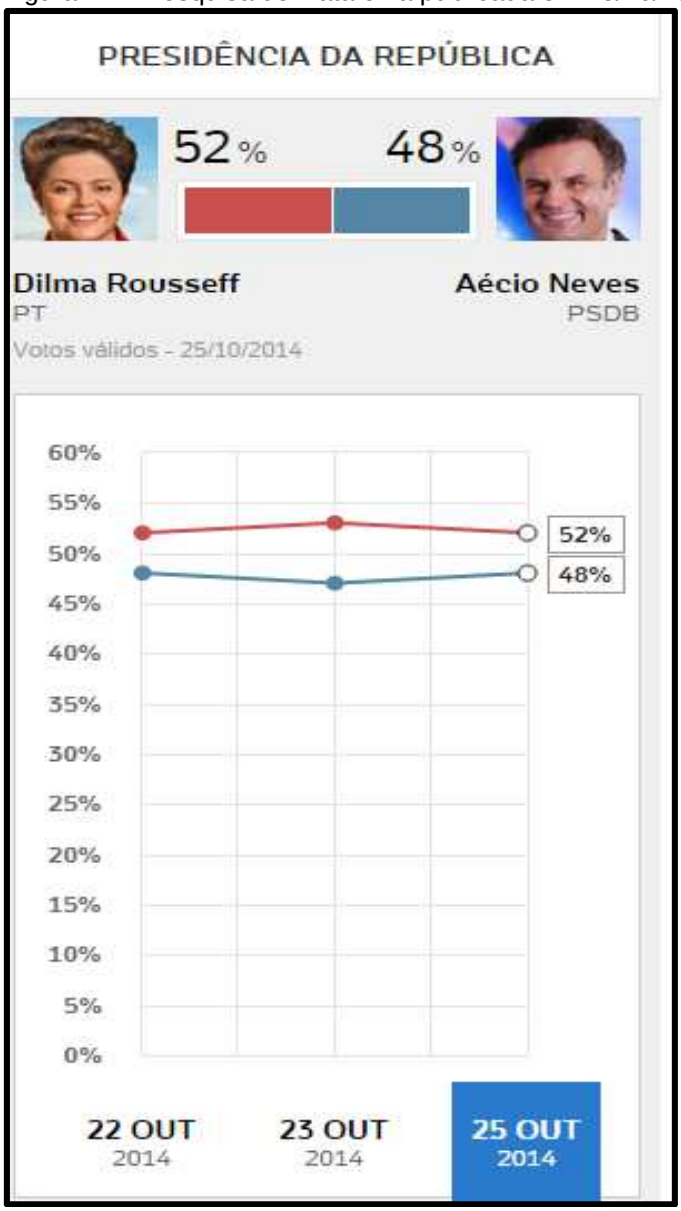
Durante a análise dos resultados, percebeu-se que se não restringisse o monitoramento com palavras-chave diretas, a busca seria ampla e apareceriam resultados que não seriam utilizados na classificação, isto ficou claro na comparação de itens encontrados nos monitoramentos. No monitoramento “Aécio Neves”, as palavras-chave “Voto em Aécio” e “Aécio Neves” formaram a base do filtro deste monitoramento, encontrando 1.713 itens, ou seja, todos os textos que contivessem qualquer uma destas duas palavras-chave estariam presentes na base dos dados, já no monitoramento “Dilma Roussef” percebeu-se que com apenas a palavra-chave “Voto em Dilma” a base de dados praticamente diminuiu pela metade o número de registros encontrados (960), sendo assim o monitoramento ganhou em rapidez na análise e na classificação dos dados.

6.2.4.1 Comparação com o Datafolha

Na comparação com o resultado da pesquisa eleitoral realizada pelo Datafolha, que foi publicada no dia 25/10/2014, data pelo qual também foi realizado a extração de dados neste trabalho, é visto que a comparação entre os monitoramentos e a pesquisa apresentou uma igualdade extrema em relação ao voto dos eleitores.

Segundo o site Uol (2014a), o Datafolha realizou a pesquisa nos dias 24/10/2014 e 25/10/2014, ouvindo 19.956 eleitores em 400 municípios. Na pesquisa, foi constatada que a candidata Dilma Roussef aparecia na primeira colocação com 52% considerando apenas os votos válidos, contra 48% do candidato Aécio Neves, o que gerou uma diferença de 4% entre um candidato e outro, conforme pode ser visto na figura 24.

Figura 24 – Pesquisa do Datafolha publicada em 25/10/2014



Fonte: Uol (2014a).

Fazendo uma comparação com o resultado dos monitoramentos, é possível chegar a conclusão que a classificação dos dados realizada neste trabalho teve coerência com relação à pesquisa feita pelo Datafolha. Conforme já visto na

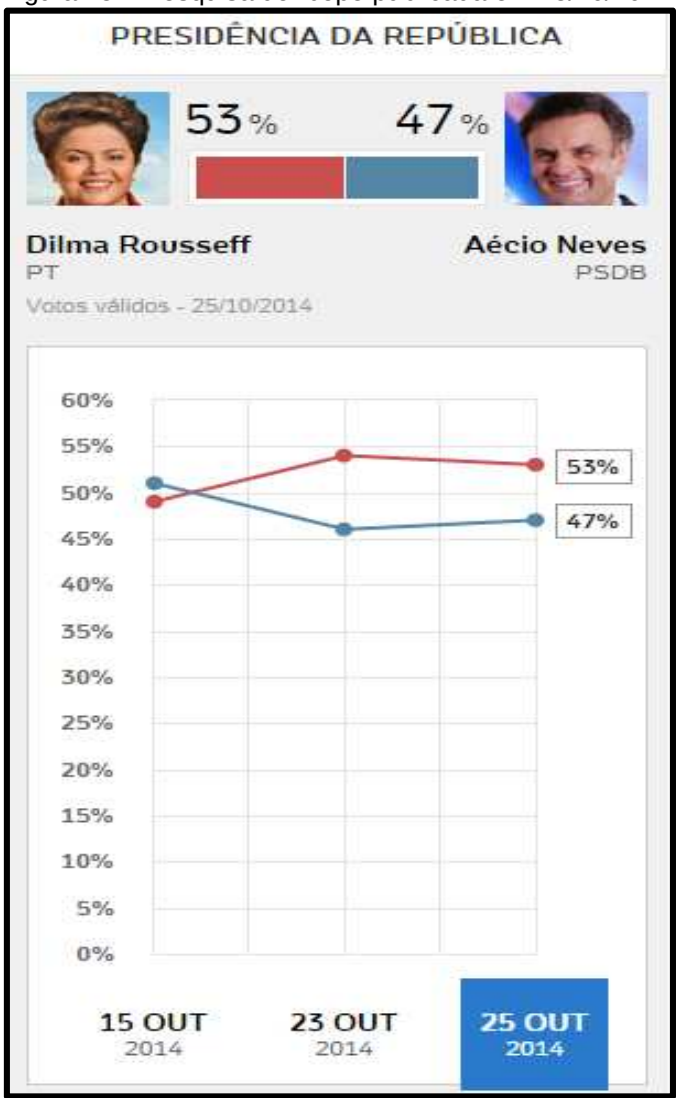
figura 21, o monitoramento “Dilma Roussef” apresentou 70,21% de menções positivas de eleitores que declaradamente votariam na atual presidente do Brasil (total de 66 menções positivas), já no monitoramento “Aécio Neves” foi visto que 66,08% (total de 113 menções positivas) do total de itens classificados escolheram o candidato à oposição Aécio Neves como seu representante no governo do país, ou seja, também existiu uma diferença de aproximadamente 4% de votos em relação aos dois candidatos comparando com a pesquisa realizada pelo Datafolha.

6.2.4.2 Comparação com o Ibope

Nas mesma condições da pesquisa realizada pelo Datafolha, segundo o site da Uol (2014a), a pesquisa do Ibope foi realizada nos dias 24/10/2014 e 25/10/2014 entrevistando 3.010 eleitores em 206 municípios e também apresentou uma igualdade se comparada à pesquisa realizada neste trabalho.

A candidata Dilma Roussef apareceu nas pesquisas em primeiro lugar com 53% dos votos válidos contra 47% do candidato Aécio Neves, gerando uma diferença de 5 pontos percentuais entre um candidato e outro, conforme pode ser visto na figura 25.

Figura 25 – Pesquisa do Ibope publicada em 25/10/2014



Fonte: Uol (2014a).

Como a comparação entre os monitoramentos realizada neste trabalho resultou em 4% de diferença entre as menções positivas dos dois candidatos conforme já visto anteriormente, é possível concluir que houve uma pequena diferença em relação à pesquisa realizada pelo Ibope, que constatou 5% de diferença entre os candidatos Dilma Roussef e Aécio Neves.

6.2.4.3 Comparação com o resultado oficial do segundo turno

Conforme dados oficiais do site Uol (2014b), a candidata Dilma Roussef se reelegeu presidente do Brasil com 51,64% dos votos (54.501.118 votos) e Aécio Neves recebeu 48,36% do total de votos (51.041.155 votos), com uma diferença de 3,4 milhões de votos, conforme pode ser visto na figura 26.

Se for considerada a diferença da porcentagem exata entre o resultado final das eleições e a comparação dos dois monitoramentos, vê-se que os resultados novamente apresentaram semelhança, pois a diferença exata da porcentagem do resultado oficial das eleições do segundo turno entre Dilma Roussef e Aécio Neves foi de 3,28% e a diferença entre a porcentagem de votos do monitoramento “Dilma Roussef” e o monitoramento “Aécio Neves” foi de 4,13%.

Figura 26 – Resultado oficial das eleições para presidente no segundo turno

Fonte: Uol (2014b).

7 CONCLUSÃO

Pode-se categorizar a mineração de dados como um importante meio de descoberta de conhecimento em bases de dados. A extração de conhecimento e a posterior análise que se pode fazer a partir de uma classificação de dados tem muita utilidade em diversas áreas.

Após o processo de classificação dos dados, foi possível evidenciar que a boa parte dos dados extraídos não foi utilizado para o estudo, concluindo então que uma pesquisa menos abrangente poderia trazer o mesmo resultado, isto pôde ser visto na comparação entre os monitoramentos, já que o monitoramento “Aécio Neves” teve praticamente o dobro de registros do monitoramento “Dilma Roussef”, pois foi utilizado uma palavra-chave a mais que trouxe muitas menções genéricas que não seriam utilizadas na classificação.

Foi visto também que muitas rotinas utilizadas nos dias de hoje podem ser auxiliadas pela mineração de dados em redes sociais, pois conforme visto nos resultados finais do desenvolvimento, a comparação da pesquisa realizada com os veículos de pesquisa Ibope e Datafolha tiveram resultados próximos se comparados com as menções positivas classificadas dos dois monitoramentos, além de se aproximar também com o resultado final das eleições presidenciais do Brasil.

Ao final do projeto, conclui-se então que a análise de opinião em redes sociais pode oferecer bons resultados a um custo relativamente baixo em relação a outros métodos utilizados.

Foi possível aplicar a mineração de dados na rede social *Twitter* utilizando a ferramenta Scup conforme programado nos objetivos do trabalho, além disso a ferramenta também serviu para a análise de toda a informação obtida, por meio de relatórios e figuras que exemplificaram e compararam os resultados finais do projeto.

Como sugestão para trabalhos futuros, considerando o projeto de pesquisa realizado, lista-se os seguintes tópicos:

a) desenvolvimento de uma ferramenta de mineração de dados especificamente para minerar em redes sociais;
b) aplicar o MSV ou outros tipos de algoritmos para mineração de dados em redes sociais em uma base de dados maior e em outra ferramenta;

c) acrescentar novas funções na ferramenta, como gravação dos dados em um banco de dados;
d) realizar a pesquisa utilizando outros filtros, monitoramentos e palavras-chave.

REFERÊNCIAS

AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun. **Mining Association Rules Between Sets of Itens in Large Databases**. ACM SIGMOD Conference Management of Data, 1993.

ALCANTARA, Mariana da Silva. **Mineração de Dados em Redes Sociais**. 2012. Monografia (Curso de Bacharelado em Sistemas de Informação). Faculdade Sete de Setembro – FASETE. Paulo Afonso (BA).

BARABÁSI, Albert-László. **How Everything is Connected to Everything else and what it means for Business, Science and Everyday Life**. Cambridge: Plume, 2003.

BARRETO, Jorge M. **Introdução às Redes Neurais.** 2002. Disponível em: < http://www.inf.ufsc.br/~barreto/tutoriais/Survey.pdf >. Acesso em: 08 jun. 2014.

BASGALUPP, Márcio Porto. **LEGAL-Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão**. Tese de Doutorado, ICMC-USP, São Carlos, 2010.

BATISTA, Alexandre da Costa. **Análise da Aplicação de Algoritmos de Data Mining em Bases de Dados de Vendas de Produtos**. Monografia. Escola Politécnica de Pernambuco, Universidade de Pernambuco, 2009.

BATISTA, Gustavo Enrique de Almeida Prado Alves. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese (Doutorado) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/>. Acesso em: 02 jun. 2014.

BENEVENUTO, Fabrício; ALMEIDA, Jussara; SILVA, Altigran. **Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações.** 2011. Disponível em: <http://homepages.dcc.ufmg.br/~fabricio/download/jai2011.pdf> Acesso em: 03 jun. 2014.

BOENTE, Alfredo Nazareno. **Descoberta de Conhecimento em Bases de Dados**. Tese de Doutorado – Departamento de Informática, AWU – American World University, 2006.

BREIMAN, Leo. et al. **Classification and Regression Trees**. Wadsworth, 1984.

BRUHA, Ivan; FAMILI, Fazel A. **Postprocessing in machine learning and data mining**. ACM, 2000.

CAMPBELL, Colin; YING, Yiming. **Learning with Support Vector Machines.** Morgan & Claypool Publishers, 2011.

CAMPOS, Kelvin Garcia. **Proposta de uma Ferramenta Para Aopio ao Processo de KDD Baseada em Técnicas de Associação**. 2008. Disponível em: <http://www.urisan.tche.br/~portalcomp/TCCs/Garcia.doc>. Acesso em: 04 jun. 2014.

CRISTIANINI, Nello; SHAWE-TAYLOR, John. **An introduction to Support Vector Machines and other Kernel-based learning methods.** Cambridge University Press, 2000.

DEGENNE, Alain; FORSÉ, Michel. **Introducing Social Networks**. London: Sage, 1999.

FACEBOOK. **Facebook Reports First Quarter 2014 Results.** 2014. Disponível em: <https://newsroom.fb.com/news/2014/04/facebook-reports-first-quarter-2014-results/>. Acesso em: 04 jun. 2014.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory.; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery: An Overview. Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, 1996.

FONSECA, Joaquim. **Indução de árvores de decisão**. Tese de Mestrado, Lisboa, 1994.

FREITAS, Alex A. **On Rule Interestingness Measures.** Knowledge Based Systems Journal, 1999.

GARTON, Laura; HAYTHORNTHWAITE, Caroline; WELLMAN, Barry. **Studying Online Social Networks. Journal of Computer Mediated Communication**. 1997. Disponível em: <http://www.ascusc.org/jcmc/vol3/issue1/garton.html>. Acesso em: 05 jun. 2014.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data Mining: um guia prático: conceitos, técnicas, ferramentas, orientações e aplicações.** Rio de Janeiro: Elsevier, 2005

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufamann Publishers, 2001.

KOHAGURA, Tiago. **Lógica Fuzzy e suas aplicações**. Trabalho de Conclusão de Curso. Universidade Estadual de Londrina, Londrina, 2007.

LAINE, Mikko; FRUHWIRTH, Christian: **Monitoring Social Media: Tools, Characteristics and Implications.** ICSOB, 2010.

LESKOVEC, Jurij; ADAMIC Lada; HUBERMAN Bernardo. **The Dynamics of Viral Marketing.** ACM Transactions on the Web, 2007.

LIBRALAO, Giampaolo Luiz et al. Determinação de vícios refrativos oculares utilizando Support Vector Machines. **Sba Controle & Automação**, Campinas, v. 16, n. 2, jun. 2005. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-17592005000200004&lng=en&nrm=iso>. Acesso em: 06 jun. 2014.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. **Uma Introdução à Support Vector Machines**. Porto Alegre, v. 14, n. 2 , p.43-67, 2007. Disponível em: <seer.ufrgs.br/rita/article/download/rita_v14_n2_p43-67/3543>. Acesso em: 06 jun. 2014.

LUCAS, Diogo C. **Algoritmos Genéticos: Uma Introdução**. Disponível em: <http://www.inf.ufrgs.br/~alvares/INF01048IA/MaterialProva2/ApostilaAlgoritmosGeneticos.pdf>. Acesso em: 07 jun. 2014.

MAICHAKI, Marcos Fernando; OLIVEIRA, Robson Leal de. **Monitoramento de Mídias Sociais.** 2012. Disponível em: <http://my-project socialmonitor.googlecode.com/svn/Artigo/modelos/Artigo - Monitoramento de Mídias Sociais.pdf>. Acesso em: 07 jun. 2014.

MILREU, Paulo. **Como monitorar o que estão falando da minha empresa**. 2010. Disponível em: <http://www.slideshare.net/milreu/como-monitorar-o-que-esto-falando-da-minhaempresa- nas-mdias-sociais-de-forma-simples-e-direta-oficina-reclame-aqui-ago2011> Acessado em: 08 jun. 2014

MIRANDA, Márcio N. D. **Algoritmos Genéticos: Fundamentos e Aplicações**. Disponível em: <http://www.gta.ufrj.br/~marcio/genetic.html>. Acesso em: 08 jun. 2014.

MOORE, Andrew W. **Instance-based learning**. School of Computer Science, Carnegie Mellon University, US, 2005. Disponível em: <http://www.autonlab.org/tutorials/mbl08.pdf>. Acesso em: 09 jun. 2014.

NIELSEN. **Pesquisa Global sobre Mídias Sociais.** 2012a.Disponível em: <http://www.nielsen.com/br/pt/insights/reports/2012/pesquisa-global-sobre-midias-sociais.html>. Acesso em: 13 jun. 2014.

NIELSEN. **State of the media: The social media reportq3**.2012b.Disponível em: <http://www.nielsen.com/us/en/reports/2012/state-of-the-media-the-social-media-report-2012.html>. Acesso em: 13 jun. 2014.

PINTO, Filipe Mota; SANTOS, Manuel Filipe. **Descoberta de Conhecimento em Bases de Dados em Actividades de CRM**. Datagadgets, 2005.

PITONI, Rafael Moreira. **Mineração de regras de associações nos canais de informação do direito**. Trabalho de Conclusão (Graduação) – Curso Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1998.

POLITO, Marco. **Data Mining**. Trabalho apresentado na disciplina de Banco de Dados do Curso de Análise de Sistemas da UERJ – Rio de Janeiro, 1997.

PRIMO, Alex F. T. **Interação Mediada por Computador: A comunicação e a educação a distância segundo uma perspectiva sistêmico-relacional**. Tese de Doutorado. Apresentada ao Programa de Pós-Graduação em Informática na Educação em março de 2003.

QUINLAN, John Ross. **Induction of decision trees. Machine Learning**. Los Altos, California: Morgan Kaufmann Publishers, 1986.

QUINLAN, John Ross. **C4.5: Programs for Machine Learning**. Los Altos, California: Morgan Kaufmann Publishers, 1993.

RECUERO, Raquel: **Redes Sociais na Internet**. Porto Alegre, Sulina, 2009.

REID, Elizabeth M. **Electropolis: Communication and Community on Internet Relay Chat**. Honoris Thesis. University of Melbourne, 1991.

RODRIGUES, Elaine. **Redes sociais: não complique, comunique-se**. 2011. Disponível em: <http://issuu.com/gaiacreative/docs/inovadoresespm_redes_sociais_inovacao_digital_vl>. Acesso em: 14 jun. 2014.

SALUSTIANO, Sérgio. **Monitoramento de Redes Sociais**. 2010. Disponível em: <www.slideshare.net/skrol/monitoramento-de-redes-sociais> Acessado em: 15 jun. 2014

SANTOS, F. C. **Variações do Método kNN e suas aplicações na Classificação Automática de Textos**. 2009. 96 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Goiás, Goiânia, GO, 2009.

SASSI, Renato José. **Uma Arquitetura Híbrida para Descoberta de Conhecimento em Bases de Dados: Teoria dos Rough Sets e Redes Neurais Artificiais Mapas Auto-Organizáveis**. Tese de Doutorado. Escola Técnica da Universidade de São Paulo, 2006.

SCUP. **Monitoramento e Gestão de Redes Sociais.**2014.Disponível em: <http://www.scup.com>. Acesso em: 15 jun. 2014.

SHAW, Ian.; SIMÕES, Marcelo Godoy. **Controle e Modelagem Fuzzy**. São Paulo: Editora Edgard Blücher, 2002.

SILVA, Tarcízio. **Para entender o Monitoramento de Mídias Sociais**. Bookes, 2012.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining - Mineração de Dados.**Ciência Moderna, 2009.

TWITTER. **Twitter Usage.** 2014. Disponível em: https://about.twitter.com/company>. Acesso em: 10 nov. 2014.

UOL (São Paulo). **Dilma é reeleita na disputa mais apertada da história; PT ganha 4º mandato.** 2014b. Disponível em: <http://eleicoes.uol.com.br/2014/noticias/2014/10/26/dilma-cresce-na-reta-final-e-reeleita-e-emplaca-quarto-mandato-do-pt.htm>. Acesso em: 29 out. 2014

UOL (São Paulo). **Ibope mostra Dilma reeleita, Datafolha aponta empate técnico.** 2014a. Disponível em:

<http://eleicoes.uol.com.br/2014/noticias/2014/10/25/ibope-mostra-dilma-reeleita-datafolha-aponta-empate-tecnico.htm>. Acesso em: 29 out. 2014.

VAPNIK, Vladimir; CHAPELLE, Olivier. **Bounds on error expectation for support vector machines**. Neural computation, 1999.

VENANCIO, Carlos Alberto Silva. **Redes Sociais: Vantagens e Desvantagens para as Empresas**. 2010. Disponível em: <http://www.mundowsionline.com/_blog/Internet_Marketing_Success_Blog/post/Redes_Sociais_-_Vantagens_e_Desvantagens_para_as_Empresas/>. Acesso em: 16 jun. 2014.

VENTURA, André Rodrigues. **Análise da Heterogeneidade Intra-tumoral em Imagens PET-FDG.** 2010. 55 f. Dissertação (Mestrado) - Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Coimbra, 2010. Disponível em: <https://estudogeral.sib.uc.pt/bitstream/10316/13813/1/Tese%20MIEB_Andr%C3%A9%20Ventura.pdf>. Acesso em: 16 jun. 2014.

WANG, Lipo. **Support Vector Machines: Theory and Applications.** Springer, 2005.

WANG, Li-Xin; MENDEL, Jerry M. **Generating fuzzy rules by learning from examples**. IEE Trans. Systems Man and Cybernetics, 1992.

WASSERMAN, Stanley; FAUST, Katherine. **Social Network Analysis. Methods and Applications**. Cambridge, UK: Cambridge University Press, 1994.

WATTS, Duncan J. **Six Degrees: The Science of a Connected Age**. New York: W. W. Norton & Company, 2003.

WATZLAWICK, Paul; BEAVIN, Janet Helmick; JACKSON, Don D. **Pragmática da Comunicação Humana**. 11ª ed. São Paulo: Cultrix, 2000.

WELLMAN, Barry. **Networks in the Global Village**. Boulder, CO: Westview Press,1999.

WHESTPHAL, Christopher; BLAXTON, Teresa. **Data Mining Solutions: Methods and Tools for Solving Real – World Problems**. John Wiley & Sons, 1998.

WINANDY, Charles-Edouard; FILHO, Estillac Borges; BENTO, Lisânia Vieira. **Algoritmos para Aprendizagem Supervisionada**. 2007. Disponível em: <http://winandy.voila.net/IA_ApSuperv_NotasAula.pdf>. Acesso em: 17 jun. 2014.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data Mining: Practical Machine Learning Tools and Techniques**. Elsevier, 2011.

YOUTUBE. **Youtube Statistics.** 2014. Disponível em: <https://www.youtube.com/yt/press/pt-BR/statistics.html>. Acesso em: 17 jun. 2014.

APÊNDICE(S)

APÊNDICE A – Artigo Científico

Aplicação da Ferramenta Scup para Análise de informação e Mineração de Dados em Redes Sociais (Twitter)

Lucas V. Cardoso[1], Paulo João Martins[1], Merisandra C. de Mattos Garcia[1]

[1]Universidade do Extremo Sul Catarinense (UNESC) – Criciúma, SC - Brasil

lucas_valdati@hotmail.com, {pjm,mem}@unesc.net

Abstract. *Social networks are being used as a data source to several studies, it is for business, public service, and marketing in various areas of computing. To perform this research, companies are using a technique called data mining, consisting of a search and sort large amounts of data in order to understand the consumer and market trends. In the case of this work, Scup tool was chosen to perform the entire data mining process, from pre-processing to final analysis of the results, it uses an algorithm based on the technique of Support Vector Machine. The research of this study was to analyze the citations related to presidential candidates of Brazil in the 2014 elections in the second round.*

Resumo. *As redes sociais estão sendo usadas como fonte de dados de várias pesquisas, seja de negócios, utilidade pública, marketing e nas diversas áreas da computação. Para realização destas pesquisas, as empresas estão utilizando uma técnica chamada mineração de dados, que consiste em uma busca e classificação de grandes quantidades de dados, afim de entender o consumidor e as tendências de mercado. No caso deste trabalho, a ferramenta Scup foi a escolhida para realizar todo o processo de mineração de dados, desde o pré-processamento até a análise final dos resultados, ela utiliza um algoritmo baseado na técnica de Máquina de Suporte Vetorial. A pesquisa deste trabalho teve o objetivo de analisar as menções relacionadas aos candidatos à presidência do Brasil nas eleições de 2014 no segundo turno.*

1. Introdução

Redes sociais online têm se tornado extremamente populares, levando ao surgimento e à crescente popularização de uma nova era de aplicações na Web. Associado a esse crescimento, redes sociais estão se tornando um tema central de pesquisas em diversas áreas da Ciência da Computação [BENEVENUTO; ALMEIDA; SILVA, 2011].

A comunicação nas redes sociais surgiu da necessidade que o ser humano tem em compartilhar com o outro, em criar laços que são construídos por afinidades existentes entre eles. Essas afinidades são assuntos em comum, ideais, preferências. Dentre diversos exemplos de assuntos em comum, podemos citar: clubes de futebol, religiões, empresas e músicas. Em meio a tantas formas de interação social, quando essa parte para o ambiente online, é que temos as chamadas redes sociais [ALCANTARA, 2012]. Atualmente, as redes sociais são as

principais formas de comunicação entre as pessoas no ambiente web, essa necessidade de compartilhar informações entre as pessoas pela Internet desencadeou um aumento significativo de informações.

Devido a esse aumento de informações e a constante presença dos internautas nas redes sociais para expressar suas opiniões sobre produtos, marcas, pessoas, gostos, costumes e assuntos corriqueiros, formou-se um grande interesse por parte de empresas e pessoas em analisar essas informações que podem ser úteis em suas estratégias de Marketing [ALCANTARA, 2012].

A mineração de dados possibilita a busca em grandes bases de dados de informações desconhecidas, permitindo aos gestores uma maior agilidade nas tomadas de decisões. Uma empresa que utiliza Data Mining é capaz de criar parâmetros para entender o comportamento do consumidor, identificando afinidades entre as escolhas de produtos e serviços, prevendo hábitos de compras e analisando comportamentos habituais para detecção de fraudes [PINTO; SANTOS, 2005].

O KDD pode ser visto como o processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração, consistindo na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis, produz uma relação particular de padrões a partir de dados [FAYYAD; PIATETSKY; SMYTH, 1996, tradução nossa].

A informação e o conhecimento obtidos podem ser utilizados para diversas aplicações, que vão do gerenciamento de negócios, controle de produção e análise de mercado ao projeto de engenharia e exploração científica [HAN; KAMBER, 2001, tradução nossa].

Desta forma, foi possível analisar informações e minerar dados nas redes sociais utilizando a base de dados do *Twitter* utilizando a ferramenta Scup.

2. Descoberta de Conhecimento em Base de Dados

O *Knowledge Discovery in Data Base* (KDD) é designado como o processo de extração de conhecimento pela análise de grandes conjuntos de dados, sua principal etapa é o processo de mineração dos dados, que consiste na ação prática de análise e de determinados algoritmos que produzem uma relação de padrões a partir de dados [FAYYAD; PIATETSKY; SMYTH, 1996, tradução nossa].

Para desenvolver o processo KDD, segundo Goldschmidt e Passos [2005], devem ser analisados três aspectos: o problema para aplicar ao KDD (Conjunto de Dados, Especialista e Objetivos), os recursos disponibilizados para atender a solicitação (Especialista KDD, Ferramentas e Hardware) e o resultado aplicado ao problema (Modelo do Conhecimento e Histórico do Conhecimento). Ainda, conforme o mesmo autor, a Descoberta de Conhecimento em base de dados é composta por três etapas: Pré-processamento, Mineração de Dados e Pós-processamento.

O objetivo do Pré-processamento está relacionado à captação, à organização e ao tratamento dos dados, ou seja, nesta etapa devemos preparar os dados para posteriormente fazer a Mineração dos Dados.

Na etapa de Mineração de dados, é feita a aplicação de algoritmos sobre os dados para obtenção do conhecimento, como também são definidas as técnicas e os algoritmos que serão utilizados. Entre as técnicas existentes, podemos citar: Redes Neurais, Algoritmos Genéticos, Modelos Estatísticos e Probabilísticos [GOLDSCHMIDT; PASSOS, 2005].

O pós-processamento compreende o tratamento do conhecimento que foi adquirido na Mineração de Dados, seu objetivo é realizar se existe utilidade no conhecimento descoberto [FAYYAD; PIATETSKY; SMYTH , 1996, tradução nossa].

2.1 Máquina de Suporte Vetorial

A Máquina de Suporte Vetorial é uma técnica de aprendizagem de máquina que cria funções para treinar e classificar os dados [WANG, 2005, tradução nossa]. O principal objetivo é obter um equilíbrio entre os erros de treinamento e ao conjunto de teste, minimizando em demasia os ajustes com as amostras de treinamento, aumentando a capacidade de generalização dos dados [VAPNIK, 1999, tradução nossa].

Segundo Campbell e Yiming [2011, tradução nossa], a MSV é uma máquina de aprendizado abstrata, onde é utilizado um conjuntos de dados para treinar e logo após o treinamento, deve ser feito a geração dos dados e fazer uma previsão sem erros do comportamento do restante dos dados.

Na figura 1 é possível exemplificar como o MSV trabalha para buscar os dados. No exemplo retas que separam os dois conjuntos de dados utilizados, ele tentará traçar uma reta que separe os dados da forma mais correta, afim de buscar os dados que mais se aproximem com os filtros utilizados. A reta que é considerada ótima é a que possui a maior capacidade de separar os dados desconhecidos.

Fonte: Santos(2009)

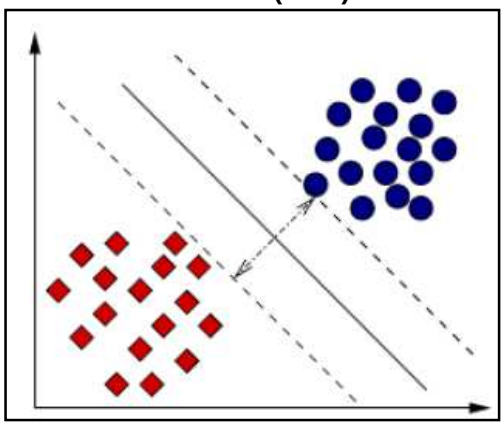


Figura 1 – Exemplo de classificador MSV

No trabalho desenvolvido, foi utilizado um algoritmo para busca e classificação de dados que se baseia no algoritmo MSV, através da ferramenta Scup.

3. Redes Sociais

As redes sociais são estruturadas em um conjugado de dois elementos, que são eles: os atores (pessoas conectadas, grupos) que são os “nós” em uma rede, e as suas conexões (ligações e laços de amizade) [WASSERMAN; FAUST,1994, tradução nossa].

Segundo o relatório do primeiro trimestre de 2014 do *Facebook* [FACEBOOK, 2014, tradução nossa], a rede social tem em sua base mais de 802 milhões de usuários ativos em sua rede. A popularidade desta rede social, está diretamente ligada as funções que o usuário pode fazer, desde publicações simples de texto até compartilhamento de vídeos e fotos, entre diversas outras ações. Segundo as estatísticas do *YouTube* [YOUTUBE, 2014], mais de um bilhão de usuários visitam o *Youtube* todos os meses, como também mais de seis bilhões de horas são assistidas por mês, que é o equivalente a uma hora para cada pessoa em

todo o mundo. No *Twitter* existem cerca de 284 milhões de usuários ativos por mês e sua principal característica é a postagem de *tweets* de no máximo 140 caracteres de texto puro [TWITTER, 2014].

Segundo a Nielsen [2012], trinta por cento dos brasileiros acessam pelo menos uma vez ao dia as redes sociais para interagir sobre marcas, produtos e serviços, sendo que 16% deles acessam mais de uma vez ao dia e 14% uma única vez.

4. Monitoramento e Extração de Conhecimento em Redes Sociais

O monitoramento de mídias sociais abrange tanto o mercado da comunicação digital quanto as ferramentas e técnicas que são utilizadas por empresas para classificar as informações encontradas. Nos últimos anos houve um surgimento de várias práticas do monitoramento de mídias sociais: empresas especializadas na prática do monitoramento foram criadas, empresas contrataram serviços de monitoramento, lançamento de publicações acerca do assunto e especializações [SILVA, 2012].

Para monitorar mídias sociais, devemos seguir os seguintes passos: coletar previamente os dados, armazenar os dados que foram coletados, classificar, categorizar, adicionar informações e analisar as menções online dos termos definidos, conforme mostra a figura 2 [SILVA, 2012].

Fonte: Silva (2012)

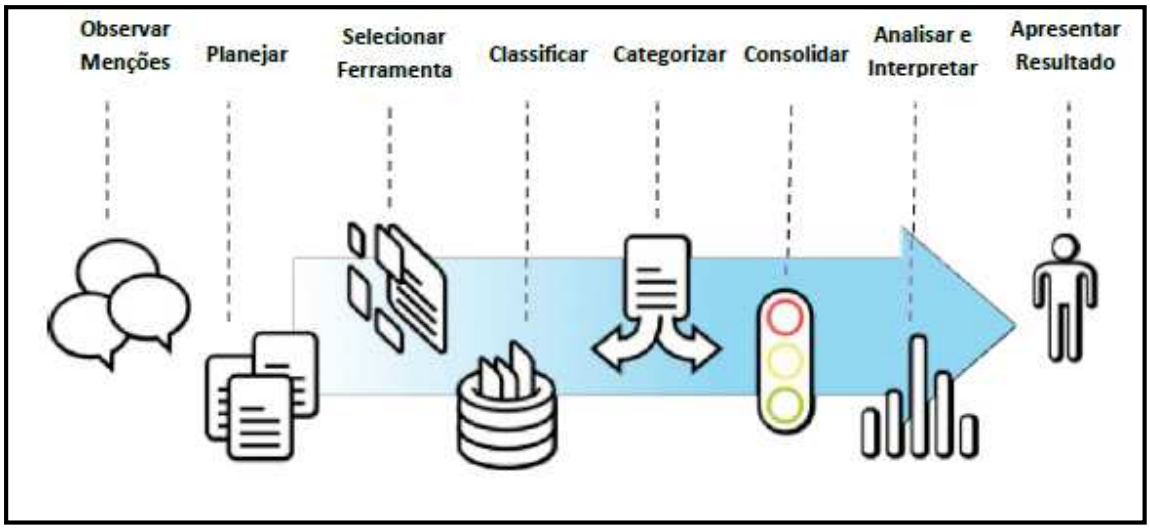


Figura 2 – Como Monitorar Mídias Sociais

Ainda segundo o mesmo autor, o principal objetivo de monitoramento das mídias sociais é a análise da opinião dos consumidores e clientes de um determinado produto ou empresa, dessa forma consegue-se conhecer melhor os públicos alvos, para posteriormente realizar ações que possam satisfazer os clientes e alcançar os objetivos das organizações.

Segundo Salustiano [2010], o monitoramento procura entender o sentimento do consumidor.

O monitoramento faz com que a organização conheça melhor os seus clientes, vigie o mercado de forma profunda, identifica novas tendências, evita denegrir a imagem da empresa e com a análise feita pode conseguir aperfeiçoar e criar novos produtos [MILREU, 2010].

4.1 Ferramenta Scup

Criado em 2009, o Scup é uma ferramenta de monitoramento pleno de mineração de conteúdos em redes sociais, segundo a própria Scup [2014] ela é uma das ferramentas mais utilizadas no Brasil (mais de 22.000 marcas). Ela coleta dados, categoriza e analisa as palavras-chave, que os analistas definem em determinadas mídias sociais. Abaixo apresenta as cadastradas [MAICHAKI; OLIVEIRA, 2012]:

a) Google;
b) Twitter;
c) Facebook;
d) Youtube;
e) Flickr;
f) Yahoo;
g) Reclame Aqui;
h) Feed Wordpress;
i) RSS;
j) Slide Share;
k) Linkedin;
l) Vimeo;
m) Tumbrl;
n) Foursquare;
o) Instagram.

A ferramenta monitora em tempo real os dados que foram colocados na pesquisa e após a finalização da mesma, é possível criar relatórios e exportá-los para tabelas e gráficos escolhendo diferentes formas [SCUP, 2014].

Ele utiliza *API's* para a coleta de registros, além de utilizar um algoritmo proveniente do método de SVM para mineração dos dados [MAICHAKI; OLIVEIRA, 2012].

5. Trabalho Desenvolvido

Com base no estudo realizado, foi aplicada a mineração de dados em redes sociais utilizando a ferramenta Scup.

Para a coleta dos dados, criou-se os monitoramentos "Aécio Neves" com as palavras-chave "Aécio Neves" e "Voto em Aécio", e o outro monitoramento "Dilma Roussef", relacionando com a palavra-chave "Voto em Dilma", afim de fazer uma comparação das menções encontradas dos usuários da rede social *Twitter* sobre quem o usuário iria votar no dia 26/10/2014, data da eleição presidencial do Brasil no 2º turno. Por existir muitas menções, os dados foram extraídos apenas no dia 25/10/2014, data que antecedia às eleições.

Mesmo com o uso de filtros para a extração dos dados monitorados, muitos *tweets* trazem informações irrelevantes para a desenvoltura do trabalho. Um exemplo disso é a postagem: "e a passeata do aécio neves novamente embaixo de minha janela". Esta postagem foi armazenada devido ao filtro "Aécio Neves" estar sendo utilizado no monitoramento criado, porém este *tweet* não apresenta informação relevante para entrar como parâmetro na comparação dos resultados. No entanto, se verificar o filtro inserido em outro texto, pode ver que ele terá relevância: "Eu vou de Aécio Neves", conforme pode ser visto na figura 3. É

perceptível que no texto o usuário declarou seu voto em Aécio Neves, ou seja, esta informação é relevante para uma futura análise de opinião. Tendo em vista estes argumentos, é indispensável a realização de uma limpeza sobre a base de dados extraída para que não interfira no resultado final.

Fonte: Scup(2014)

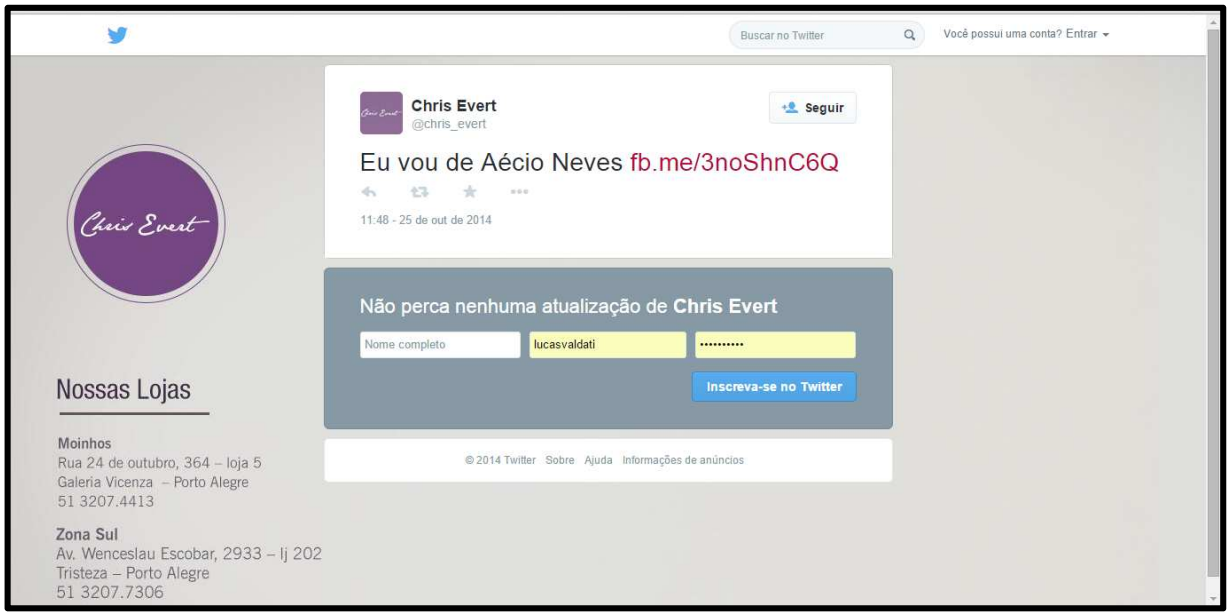


Figura 3 – Exemplo de menção encontrada na base de dados

Na etapa de mineração dos dados, ocorre a extração de informação dos dados. Para execução deste processo serão aplicadas técnicas baseadas no algoritmo MSV, disponibilizadas na própria ferramenta, que têm como objetivo à classificação das menções em positivo, negativo e neutro em relação aos candidatos à presidência. Tais dados encontrados por meio dos filtros, fazendo a execução da MSV, ou seja, cada filtro ou palavra utilizada como filtro representa uma dimensão e esta é separada das demais pela utilização do algoritmo empregado.

Por fim, todo o processo executado anteriormente irá ser despontado e comparado por meio de gráficos e tabelas que representam a classificação das menções encontradas na base de dados. Os dados classificados podem ser comparados aos resultados, em alguns veículos de pesquisa do país e propriamente ao resultado oficial das eleições no 2º turno.

O objetivo final é verificar se a pesquisa e a classificação executadas obtiveram uma coerência com o resultado final das eleições.

6. Resultados Obtidos

A base de dados utilizada possuía dois monitoramentos: Aécio Neves (com as palavras-chave Aécio Neves e Voto em Aécio) e Dilma Roussef (com a palavra-chave Voto em Dilma). Após a etapa do pré-processamento que visava a limpeza dos dados que não são úteis como informação, ocorreu a etapa de mineração dos dados ou classificação dos dados, para posteriormente realizar a análise de toda a informação extraída.

Com o uso da ferramenta de mineração de dados Scup, foram coletados ao todo 2.673 itens no dia 25/10/2014. Deste 1.713 *tweets* do monitoramento “Aécio Neves” e 960 *tweets* do monitoramento “Dilma Roussef”, conforme pode ser visto na figura 4.

Fonte: Scup(2014)

Figura 4 – Quantidade de itens coletados

6.1 Comparação com o Resultado Oficial das Eleições

Conforme dados oficiais do site Uol [2014], a candidata Dilma Roussef se reelegeu presidente do Brasil com 51,64% dos votos (54.501.118 votos) e Aécio Neves recebeu 48,36% do total de votos (51.041.155 votos), com uma diferença de 3,4 milhões de votos, conforme pode ser visto na figura 5.

Se for considerada a diferença da porcentagem exata entre o resultado final das eleições e a comparação dos dois monitoramentos, vê-se que os resultados novamente apresentaram semelhança, pois a diferença exata da porcentagem do resultado oficial das eleições do segundo turno entre Dilma Roussef e Aécio Neves foi de 3,28% e a diferença entre a porcentagem de votos do monitoramento "Dilma Roussef" e o monitoramento "Aécio Neves" foi de 4,13%.

Fonte: Uol (2014)

Figura 5 – Resultado oficial das eleições para presidente no segundo turno

7. Conclusão

Pode-se categorizar a mineração de dados como um importante meio de descoberta de conhecimento em bases de dados. A extração de conhecimento e a posterior análise que se pode fazer a partir de uma classificação de dados tem muita utilidade em diversas áreas.

Após o processo de classificação dos dados, foi possível evidenciar que a boa parte dos dados extraídos não foi utilizado para o estudo, concluindo então que uma pesquisa menos abrangente poderia trazer o mesmo resultado, isto pôde ser visto na comparação entre os monitoramentos, já que o monitoramento "Aécio Neves" teve praticamente o dobro de registros do monitoramento "Dilma Roussef", pois foi utilizado uma palavra-chave a mais que trouxe muitas menções genéricas que não seriam utilizadas na classificação.

Foi visto também que muitas rotinas utilizadas nos dias de hoje podem ser auxiliadas pela mineração de dados em redes sociais, pois conforme visto nos resultados finais do desenvolvimento, a comparação da pesquisa realizada com os veículos de pesquisa Ibope e Datafolha tiveram resultados próximos se comparado com as menções positivas classificadas dos dois monitoramentos, além de se aproximar também com o resultado final das eleições presidenciais do Brasil.

Ao final do projeto, conclui-se então que a análise de opinião em redes sociais pode oferecer bons resultados a um custo relativamente baixo em relação a outros métodos utilizados.

Foi possível aplicar a mineração de dados na rede social *Twitter* utilizando a ferramenta Scup conforme programado nos objetivos do trabalho, além disso a ferramenta também serviu para a análise de toda a informação obtida, por meio de relatórios e figuras que exemplificaram e compararam os resultados finais do projeto.

8. Referências

ALCANTARA, Mariana da Silva. **Mineração de Dados em Redes Sociais**. 2012. Monografia (Curso de Bacharelado em Sistemas de Informação). Faculdade Sete de Setembro – FASETE. Paulo Afonso (BA).

BENEVENUTO, Fabrício; ALMEIDA, Jussara; SILVA, Altigran. **Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações.** 2011. Disponível em: <http://homepages.dcc.ufmg.br/~fabricio/download/jai2011.pdf> Acesso em: 03 jun. 2014.

CAMPBELL, Colin; YING, Yiming. **Learning with Support Vector Machines.** Morgan & Claypool Publishers, 2011.

FACEBOOK. **Facebook Reports First Quarter 2014 Results.** 2014. Disponível em: <https://newsroom.fb.com/news/2014/04/facebook-reports-first-quarter-2014-results/>. Acesso em: 04 jun. 2014.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory.; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery: An Overview. Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, 1996.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data Mining: um guia prático: conceitos, técnicas, ferramentas, orientações e aplicações.** Rio de Janeiro: Elsevier, 2005

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufamann Publishers, 2001.

MAICHAKI, Marcos Fernando; OLIVEIRA, Robson Leal de. **Monitoramento de Mídias Sociais.** 2012. Disponível em: <http://my-project socialmonitor.googlecode.com/svn/Artigo/modelos/Artigo - Monitoramento de Mídias Sociais.pdf>. Acesso em: 07 jun. 2014.

MILREU, Paulo. **Como monitorar o que estão falando da minha empresa**. 2010. Disponível em: <http://www.slideshare.net/milreu/como-monitorar-o-que-esto-falando-da-minhaempresa-nas-mdias-sociais-de-forma-simples-e-direta-oficina-reclame-aqui-ago2011> Acessado em: 08 jun. 2014

NIELSEN. **Pesquisa Global sobre Mídias Sociais.** 2012.Disponível em: <http://www.nielsen.com/br/pt/insights/reports/2012/pesquisa-global-sobre-midias-sociais.html>. Acesso em: 13 jun. 2014.

PINTO, Filipe Mota; SANTOS, Manuel Filipe. **Descoberta de Conhecimento em Bases de Dados em Actividades de CRM**. Datagadgets, 2005.

RECUERO, Raquel: **Redes Sociais na Internet**. Porto Alegre, Sulina, 2009.

SALUSTIANO, Sérgio. **Monitoramento de Redes Sociais**. 2010. Disponível em: <www.slideshare.net/skrol/monitoramento-de-redes-sociais> Acessado em: 15 jun. 2014

SANTOS, F. C. **Variações do Método kNN e suas aplicações na Classificação Automática de Textos**. 2009. 96 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Goiás, Goiânia, GO, 2009.

SCUP. **Monitoramento e Gestão de Redes Sociais.**2014.Disponível em: <http://www.scup.com>. Acesso em: 15 jun. 2014.

SILVA, Tarcízio. **Para entender o Monitoramento de Mídias Sociais**. Bookes, 2012.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining - Mineração de Dados.**Ciência Moderna, 2009.

TWITTER. **Twitter Usage.** 2014. Disponível em: https://about.twitter.com/company>. Acesso em: 10 nov. 2014.

UOL (São Paulo). **Dilma é reeleita na disputa mais apertada da história; PT ganha 4º mandato.** 2014b. Disponível em: <http://eleicoes.uol.com.br/2014/noticias/2014/10/26/dilma-cresce-na-reta-final-e-reeleita-e-emplaca-quarto-mandato-do-pt.htm>. Acesso em: 29 out. 2014

VAPNIK, Vladimir; CHAPELLE, Olivier. **Bounds on error expectation for support vector machines**. Neural computation, 1999.

WANG, Lipo. **Support Vector Machines: Theory and Applications.** Springer, 2005.

WASSERMAN, Stanley; FAUST, Katherine. **Social Network Analysis. Methods and Applications**. Cambridge, UK: Cambridge University Press, 1994.

YOUTUBE. **Youtube Statistics.** 2014. Disponível em: <https://www.youtube.com/yt/press/pt-BR/statistics.html>. Acesso em: 17 jun. 2014.