UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM ENGENHARIA DE SOFTWARE

JOSÉ ADAIL CARVALHO FILHO

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTO UTILIZANDO TWEETS REFERENTES À COPA DO MUNDO 2014

QUIXADÁ
Novembro, 2014

JOSÉ ADAIL CARVALHO FILHO

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTO UTILIZANDO TWEETS REFERENTES À COPA DO MUNDO 2014

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel.

Área de concentração: computação

Orientadora Profa. Ticiana Linhares Coelho da Silva

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

C323m Carvalho Filho, José Adail
Mineração de textos: análise de sentimentos utilizando Tweets referentes à Copa do Mundo 2014 / José Adail Carvalho Filho. – 2014.
46 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2014.
Orientação: Profa. Me. Ticiana Linhares Coelho da Silva
Área de concentração: Computação

1. Mineração de dados (Computação) 2. Redes sociais on-line 3. Twitter (Rede social on-line) I. Título.

CDD 004.6

JOSÉ ADAIL CARVALHO FILHO

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTO UTILIZANDO TWEETS REFERENTES À COPA DO MUNDO 2014

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel.

Área de concentração: computação

Aprovado em: _____ / novembro / 2014.

BANCA EXAMINADORA

_____
Prof. MSc. Ticiana Linhares Coelho da Silva
(Orientadora)
Universidade Federal do Ceará-UFC

_____
Prof. MSc. José Moraes Feitosa
Universidade Federal do Ceará-UFC

_____
Prof. MSc. Paulo Antônio Leal Rego
Universidade Federal do Ceará-UF

Dedico este trabalho a minha família, por todo apoio me dado para chegar até aqui, e além.

AGRADECIMENTOS

Aos meus pais, Adail Carvalho e Ana Cláudia, que mesmo na simplicidade se esforçaram para me dar a melhor educação, sempre me motivando a trilhar bons caminhos.

Aos meus irmãos Darlly e Adaiana, pelo companheirismo e bagunça que me faz esquecer quaisquer aflições.

Agradeço minha tia Anis Lúcia, pelos bons conselhos, e pelo carisma sem par, que sempre torna meu dia especial.

Ao Lucas, Sérgio, Jonas, Amaro, Brendo, Baiano e Ygor, meus irmãos nessa cidade, longe da família de sangue, a quem sempre posso contar, independente do momento.

A Fernanda Cunha, pela amizade verdadeira, e por toda a paciência de me escutar e ser companheira, não importa o momento, e me estabiliza nos momentos em que os problemas ameaçam arruinar meu dia.

Agradeço especialmente a minha orientadora Ticiana Linhares, pelo conhecimento e oportunidades que adquiri como aluno dela, engrandecendo minha vida acadêmica e me motivando a ser um grande profissional.

Agradeço a todos os meus amigos e minha turma, especialmente a de Quixadá, por fazerem de minha jornada aqui inesquecível e prazerosa de recordar.

Ao anjo protetor enviado por Deus, que me permitiu estar aqui agradecendo, e encerrando um ciclo importante em minha vida.

Amo a todos. Amém.

.

“Se as coisas são inatingíveis... Ora! Não é motivo para não querê-las.
Que triste os caminhos, se não fora a mágica presença das estrelas.”
(Mário Quintana)

RESUMO

O aumento das redes sociais nos últimos anos permitiu aos usuários se conectarem e compartilharem informações em tempo real, enviando-as a milhares de outros usuários em um curto espaço de tempo. Além disso, a maneira como os usuários interagem mudou. Os usuários de redes sociais costumam postar suas opiniões sobre os grandes eventos, lançamentos de produtos, catástrofes, epidemias, entre outros acontecimentos. Para acompanhar o que eles estão falando nas redes sociais pode ser um diferencial para as organizações que desejam elaborar melhores estratégias de marketing, obter *feedback* sobre algum produto ou determinado evento. No entanto, essa grande quantidade de dados ainda continua crescendo, e a análise desses dados de forma não automatizada pode ser um problema não trivial. Neste contexto, este trabalho mostra como o processo de mineração de textos foi usado para coletar, estruturar o texto extraído do Twitter e como criar um modelo de classificação de texto que permita mapear a opinião da rede social dos usuários do Twitter sobre Copa do Mundo da FIFA Brasil 2014. As postagens dos usuários, popularmente conhecido como *tweets*, são categorizadas neste trabalho como um sentimento: positivo, negativo, ambíguo ou neutro.

Palavras-chave***:*** Classificação de textos. Mineração de opiniões. Redes sociais.

ABSTRACT

The increase of the social networks in the last years allowed users to get connected and share information in real time, spreading it for thousands of others users in a short time. Also, the way that users interact has changed. The users of social networks usually post their opinions about big events, product releases, catastrophes, epidemics, among other happenings. To follow what they are talking on the social networks may be a differential for organizations who wants to elaborate better marketing strategies, to obtain feedback about some product or a certain event, among other possibilities. Although, this big amount of data still keeps growing, and analysis it in a non-automated way may be a non-trivial problem. In this context, this article shows how the Text Mining process was used to collect, to structure the text extracted from Twitter(tweets) and to create a text classification model that allowed to predict the Twitter social network user's opinion about the FIFA World Cup Brazil 2014. The user's posts called as tweets are categorized in this work as a sentiment: positive, negative, ambiguous or neutral.

Keywords: Opinion mining. Social networks. Text classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O rápido crescimento do uso da Internet e a popularização das redes sociais mudaram a forma de interação entre pessoas e organizações. Usuários publicam suas opiniões sobre as organizações, eventos, catástrofes, dentre outros, em seus perfis em redes sociais, que de maneira rápida são propagadas para vários outros usuários. Estas mensagens podem conter teores positivos, mas também podem ser duras críticas. Isso pode acarretar em vantagens para as organizações, mas também pode comprometer seriamente a imagem das mesmas, pelo enorme alcance de usuários que atualmente as redes sociais possuem.

Essa popularização da Internet, por sua vez, gera um grande volume de informação a cada instante, e as organizações, em geral, não conseguem acompanhar no mesmo ritmo o que os usuários estão comentando sobre as mesmas. No entanto, percebeu-se que ao analisar essas informações, as organizações poderiam ter a vantagem de conhecer as opiniões dos usuários sobre seus serviços ou produtos fornecidos a partir de dados das redes sociais (Gomes, 2013).

Neste contexto, a Mineração de Textos, também conhecida como Descoberta de Conhecimento em Textos fornece um conjunto de técnicas que podem automatizar o processo de coleta e estruturação de informações e, junto com a Análise de Sentimentos, permite que as organizações possam saber o que os usuários estão comentando sobre elas em seus perfis na *web*. A partir dessa análise, as organizações podem se beneficiar dos resultados para os mais diversos fins, como elaborar estratégias de *marketing*, táticas de segurança, melhoria de serviços, dentre outros.

A Copa do Mundo ocorreu no Brasil neste ano. Ela é a maior competição internacional de esporte único disputado pelas seleções de futebol masculinas principais das 208 federações afiliadas à FIFA. Tendo em vista que esse é um evento de grande cobertura e interesse social, pela popularidade do futebol no Brasil e também cercado de polêmicas e manifestações devido ao seu grande custo, em contraste com outros problemas que a população brasileira julga mais importantes[1], este trabalho tem como meta a coleta, estruturação e descoberta de conhecimento de dados textuais extraídos do Twitter relacionados à Copa, para mapear a opinião dos usuários sobre o evento.

[1] http://jcrs.uol.com.br/site/noticia.php?codn=141091

O objetivo deste trabalho foi categorizar os *tweets* de acordo com o sentimento expresso. Além disso, a classificação dos sentimentos expressos nos *tweets* minerados foi validade com os eventos relacionados à Copa.

Alguns trabalhos, como o de Rodrigues Barbosa et al (2012), já fizeram análise de sentimentos em eventos importantes que mobilizam a população. Rodrigues Barbosa et al (2012) coletaram dados das eleições presidenciais do Brasil em 2010, utilizando *hashtags* (marcadores que agrupam *tweets*), para traçar o sentimento da população a respeito das eleições daquele ano. Neste trabalho, também foram utilizadas as *hashtags* para analisar os sentimentos expressos nos *tweets* que falam sobre a Copa do Mundo 2014, além de terem sido consideradas para a classificação palavras que não eram *hashtags* mas expressavam um determinado sentimento. Dessa forma, os *tweets* coletados foram classificados em um sentimento: positivo, negativo, ambíguo ou neutro. O resultado da classificação permitiu conhecer a opinião dos usuários do Twitter sobre a Copa.

A Seção 2 apresenta os conceitos chaves deste trabalho. Na Seção 3, são apresentados os trabalhos relacionados, utilizados como base de conhecimento para a realização deste trabalho. Na Seção 4 são apresentados os objetivos geral e específico deste trabalho. A Seção 5 mostra os passos seguidos para a execução deste trabalho. A Seção 6 apresenta os resultados deste trabalho. Na seção 7 é apresentada os trabalhos futuros. A seção 8 apresenta a conclusão deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Mineração de Dados

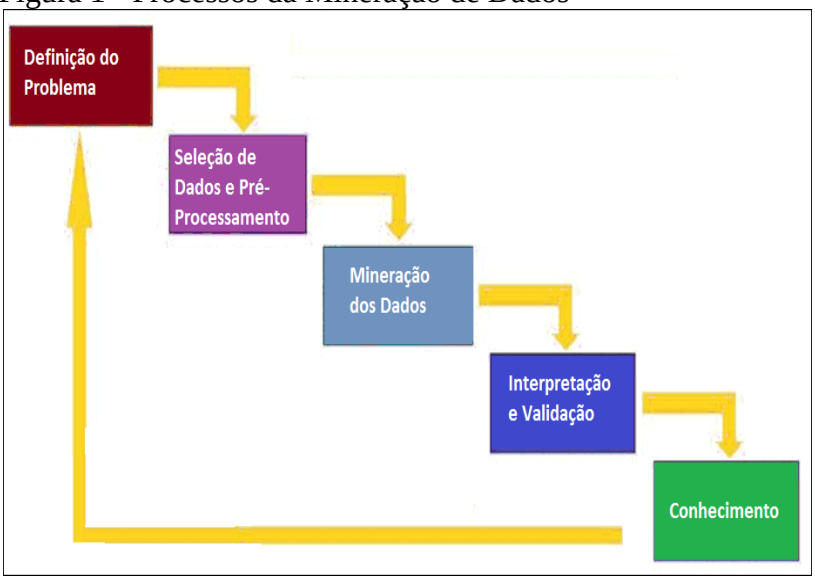
A Mineração de Textos é a principal área de estudo deste trabalho. Ela pode ser vista como uma extensão da Mineração de Dados (TAN, 1999).

A Mineração de Dados, também conhecida como Descoberta de Conhecimento em Banco de Dados, foca na exploração computadorizada em grandes massas de dados para descobrir padrões interessantes entre elas (FELDMAN et al, 1998). A maioria dos trabalhos de mineração de dados é realizada em cima de base de dados estruturada.

Segundo Morais e Ambrósio (2007), é importante que os resultados do processo de Descoberta de Conhecimento em Banco de Dados sejam compreensíveis para nós humanos, e principalmente para os usuários finais do processo, que são geralmente pessoas que tomam decisões nas organizações. Os processos de Mineração de Dados (Figura 1) devem ser vistos como práticas para melhorar os resultados das explorações feitas utilizando ferramentas tradicionais de exploração de dados.

O processo de Descoberta de Conhecimento em Banco de Dados se assemelha ao de Descoberta de Conhecimento em Textos, desde a definição do problema, até a extração de conhecimentos. Na definição do problema, os objetivos a serem alcançados através da DCBD são identificados. Na seleção de dados e pré-processamento, os dados normalmente não estão em um formato adequado para extração de conhecimento, por isso faz-se necessária a aplicação de métodos para extração e integração, transformação, limpeza, seleção e redução de volume destes dados, antes da etapa de mineração (MORAIS; AMBRÓSIO, 2007). Na fase de mineração de dados, busca-se cumprir os objetivos determinados na fase de definição de problema. É nessa etapa que são configurados e executados algoritmos em busca de padrões relevantes, podendo executá-los diversas vezes, até que sejam alcançados os resultados esperados. Na etapa de conhecimento, diversos padrões podem ser identificados, alguns não relevantes, outros interessantes para o domínio do problema.

Figura 1 - Processos da Mineração de Dados



Fonte: elaborado pelo autor.

É importante apresentar os resultados mais relevantes aos usuários. Caso não tenha obtido os resultados esperados, é possível redefinir os problemas nos quais se deseja extrair conhecimento sobre, repetindo o processo até adquirir resultados mais relevantes.

O processo de mineração de dados se assemelha ao de Mineração de Textos, porém, ele lida com dados estruturados enquanto a Mineração de Textos é aplicada sobre dados não estruturados. O conceito de Mineração de Textos é explicado em seguida.

2.2 Mineração de Textos

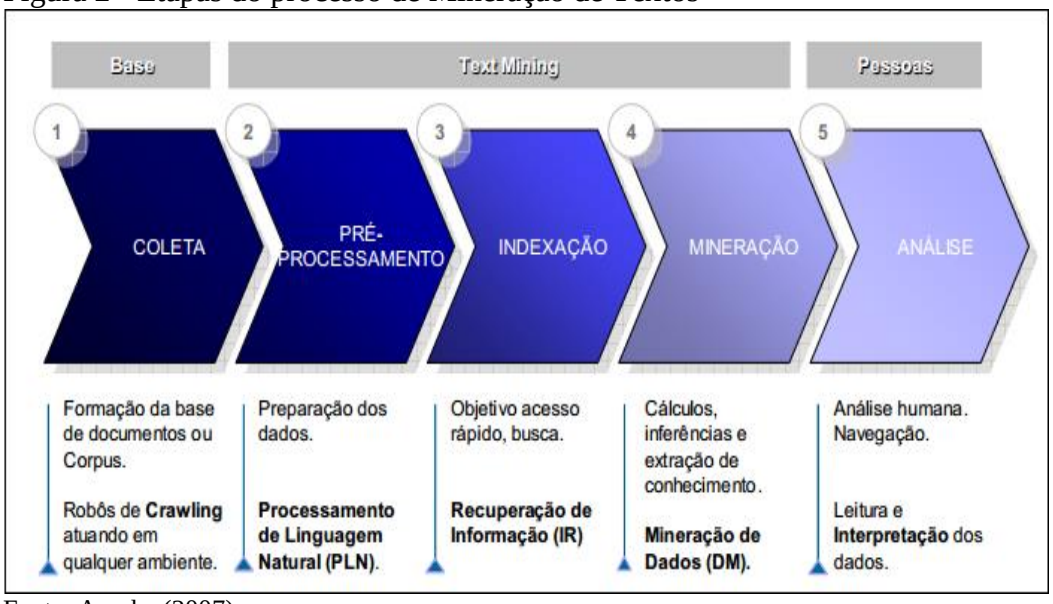
"A Mineração de Textos, também conhecida como Descoberta de Conhecimento de Texto refere-se ao processo de extrair padrões interessantes e não triviais ou conhecimento a partir de textos desestruturados." (TAN, 1999, p.1, tradução livre).

A Mineração de Textos aplica as mesmas funções analíticas da Mineração de Dados (GOMES, 2013), porém para dados textuais. Segundo Hearst (1999), os dados textuais englobam uma vasta e rica fonte de informação, mesmo em um formato que seja difícil de extrair de maneira automatizada.

Esta grande massa de informação textual não estruturada não pode ser utilizada por computadores para extração de conhecimento, uma vez que os mesmos a tratam apenas como uma sequência de caracteres. Assim, faz-se necessária a aplicação de diferentes métodos e algoritmos para dar estruturação aos dados textuais, visando facilitar a extração de conhecimento dos respectivos dados. Embora a Mineração de Textos geralmente se refira à

extração de conhecimento de base de dados textuais, pode-se recorrer a outras áreas para facilitar o processo de descoberta de conhecimento nos dados textuais.

Figura 2 - Etapas do processo de Mineração de Textos



Fonte: Aranha (2007).

Aranha (2007) propõe um modelo de Mineração de Textos (Figura 2), contendo cinco fases distintas: coleta, pré-processamento, indexação, mineração e análise. Na coleta, utiliza-se *web crawlers*, programas que visitam sítios e extraem informações, para extração dos textos que serão utilizados para a extração de conhecimento. No pré-processamento, são utilizadas técnicas como o Processamento de Linguagem Natural para estruturar os textos que serão analisados. A indexação é a etapa onde são extraídos conceitos dos documentos através da análise de seu conteúdo e traduzidos em termos da linguagem de indexação. Esta representação identifica o documento e define seus pontos de acesso para consultas (GOMES, 2006). Na etapa de mineração, são aplicados métodos e algoritmos para a identificação de padrões interessantes e extração de conhecimento. Na parte de análise, os resultados são avaliados e validados.

Os passos de execução deste trabalho são baseados no processo de Mineração de Textos proposto por Aranha (2007), com exceção do passo de indexação, pois não foi considerado relevante para a execução deste trabalho, pois não houve necessidade de realizar consultas nos textos que compõem a base de dados.

2.2.1 Processamento da Linguagem Natural

Chama-se Processamento de Linguagem Natural (PLN) um conjunto de técnicas teórico-computacionais que visam representar dados textuais e processar a linguagem natural humana para diversas tarefas.

> O Processamento da Linguagem Natural é um conjunto de técnicas computacionais para analisar e representar ocorrências naturais de texto em um ou mais níveis de análise linguística com o objetivo de se alcançar um processamento de linguagem similar ao humano para uma série de tarefas ou aplicações. (LIDDY, 2001, p.1, tradução livre).

O PLN naturalmente lida com diversos elementos linguísticos e estrutura gramatical, sendo um processo complexo, paralelo à complexidade da linguagem natural. Liddy (2001) mostra que para processar a linguagem natural, o PLN a representa em diversos níveis, como léxico, morfológico, semântico, etc.

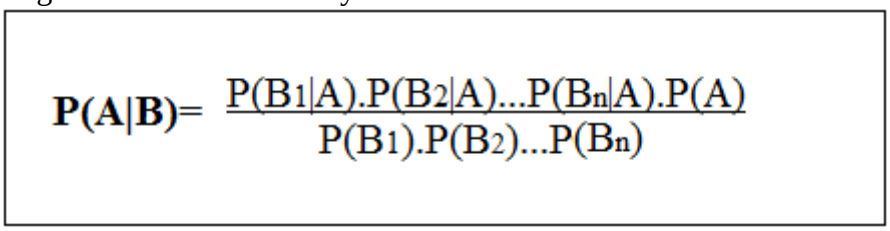
Para auxiliar na etapa de pré-processamento de dados textuais, técnicas de PLN podem ser utilizadas, como remoção de *stopwords*, segmentação de palavras, lematização, dentre outras. *Stopwords* são palavras muito comuns que aparecem no texto e carregam pouco significado; servem apenas com uma função sintática mas não indicam importância ao assunto (EL-KHAIR, 2006). Alguns exemplos de *stopwords*: as, e, os, de, para, com.

Tendo em vista a estruturação dos dados a serem utilizados nesse trabalho para melhorar o procedimento de classificação de texto, a PLN fornece técnicas que podem ser utilizadas para melhor estruturar os dados que serão utilizados neste trabalho, como por exemplo, removendo palavras dos *tweets* que serão minerados que não expressam algum sentimento ou polaridade, como artigos, preposições, dentre outros elementos gramaticais.

2.2.2 Naive Bayes

O Naive Bayes é um simples classificador probabilístico baseado na aplicação do teorema de Bayes. Ele é frequentemente utilizado como base na classificação de textos por ser rápido e fácil de implementar (RENNIE et al, 2003, p. 1, tradução livre). Gomes (2013) cita que o classificador Naive Bayes é considerado um dos mais eficientes em questões relacionadas com processamento e precisão na classificação de novas amostras.

Figura 3 - Teorema de Bayes



Fonte: elaborado pelo autor.

A Figura 3 ilustra o teorema de Bayes. Assumindo que B representa um evento que ocorreu previamente e A um evento que depende de B, para que seja calculada a probabilidade de A ocorrer dado o evento B, o algoritmo deverá contar o número de casos em que A e B ocorrem juntos e dividir pelo número de casos em que B ocorre sozinho.

O algoritmo bayesiano utilizado na etapa de Mineração de Textos neste trabalho é uma implementação pertencente ao Apache Mahout. O Mahout é uma biblioteca de aprendizagem de máquina de código aberto do Apache (OWEN,2011).

2.3 Análise de Sentimentos

A larga expansão da internet gera muitas informações em forma de opiniões em seus diversos canais: fóruns, comunidades, redes sociais, etc. Indurkhya e Damerau (2010) citam que as opiniões são tão importantes que, onde quer que se queira tomar decisões, as pessoas querem ouvir a opinião de outros. Isso não é uma verdade apenas para as pessoas, como também para as organizações, afinal, conhecer a opinião dos clientes acerca dos seus produtos e serviços é de grande valia para as organizações.

A Análise de Sentimentos ou Mineração de Opinião é o estudo computacional de opiniões, sentimentos e emoções expressos em texto. A informação textual pode ser classificada em dois principais tipos: fatos e opiniões. Fatos são expressões objetivas sobre entidades, eventos e suas propriedades. Opiniões são geralmente expressões subjetivas que descrevem os sentimentos da população, avaliações, ou sentimentos em relação às entidades, e suas propriedades (INDURKHYA; DAMERAU, 2010). Ela é a área que ajuda de maneira automatizada a determinar a direção do sentimento em textos (positivo ou negativo).

Gomes (2013) pontua que apesar da Análise de Sentimentos ser apresentada por grande parte da literatura como estudo computacional de sentimentos, a mesma pode ser utilizada para muitos outros projetos. Tendo em vista que a Análise de Sentimentos se trata de

um problema de classificação, ela pode ser utilizada para classificar dados textuais, segundo sua polaridade, mesmo se o texto não denotar algum sentimento.

2.4 Twitter

Com mais de 600 milhões de usuários cadastrados[2], o Twitter é uma rede social que permite a seus usuários postarem mensagens de texto rápidas, limitadas a 140 caracteres, conhecidas como *tweets*. Sua estrutura dinâmica permite que qualquer usuário tenha acesso às informações que são constantemente postadas, sem que para isso restrinja aos usuários possuírem alguma permissão de conexão entre eles. Russel (2013) aponta que esse é o grande diferencial do Twitter em relação a outras redes sociais populares, como o LinkedIn e o Facebook, tornando o Twitter uma rede social mais interessante de se explorar.

O Twitter fornece uma REST API para desenvolvedores, que permite aos mesmos acessarem atualizações, status, dados de usuários, etc. REST (*Representation State Transfer*) é uma arquitetura de redes de estilo híbrido derivada do estilo de várias arquiteturas baseadas em rede[3], que define uma interface conectora que permite aos clientes "conversar" com servidores de maneira única. O Twitter também fornece aos desenvolvedores acesso a um grande volume de informações em tempo real através de uma *Streaming* API. A Twitter *Streaming* API fornece para desenvolvedores acessos de baixa latência para a *stream* global do Twitter de dados de *tweets*[4].

Os *tweets* podem ser agrupados por *hashtags* palavras precedidas pelo caractere #, utilizado para marcar palavras-chave ou tópicos em um *tweet.*[5] Além disso, os usuários podem dar *retweet* em um *tweet* publicado por outros usuários. Um *retweet* é uma nova postagem do *tweet* de alguém[6]. O autor deste trabalho manteve os *retweets* para análise, por entender que quando um usuário "retuita" um *tweet* de alguém, ele está concordando com a opinião de quem postou o *tweet*.

O Twitter foi utilizado como fonte de exploração de dados para este trabalho, tendo em vista que o Twitter é uma rede social com uma grande quantidade de usuários ativos e possui alcance global, que induz seus usuários a compartilhar suas ideias constantemente,

[2] http://www.statisticbrain.com/twitter-statistics/
[3] http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
[4] https://dev.twitter.com/docs/api/streaming
[5] http://support.twitter.com/articles/255508-o-que-sao-os-marcadores-simbolos-de
[6] https://support.twitter.com/articles/263102-perguntas-frequentes-sobre-retweets-rts

gerando grande quantidade de informação a cada instante, além de fornecer APIs que permitem explorar dados dos *tweets* de maneira fácil.

3 TRABALHOS RELACIONADOS

3.1 Mineração de Textos

Tan (1999) e Gomes (2013) pontuam a relevância da aplicação de métodos de Mineração de Textos para a extração de conhecimento em base de dados textuais, uma vez que a Mineração de Dados é comumente aplicada em dados que possuem certo nível de estruturação e contemplam apenas uma parte limitada de dados que as organizações possuem, ou seja, dados estruturados (GOMES, 2013).

Gomes (2013) aplica a Mineração de Textos em seu trabalho na busca de extrair conhecimento acerca de notícias de economia de Portugal. Assim, o autor visita sítios de notícias sobre economia de seu país para representar o sentimento expresso nas notícias, analisando os textos dos títulos das notícias.

Rodrigues Babosa et al (2012) utilizam os processos da Mineração de Textos para explorar *tweets* que falam sobre as eleições presidenciais do Brasil do ano de 2010, afim de traçar o sentimento online da população expresso nos *tweets*, classificando-os em positivos, negativos, neutros ou ambíguos, e correlacionar a classificação dos *tweets* aos fatos que ocorriam no Brasil relacionados às eleições, como debates políticos, por exemplo.

Neste trabalho, será aplicado o processo de Mineração de Textos semelhante ao trabalho de Gomes (2013), no entanto a aplicação será na rede social Twitter. A escolha de usar essa rede social é pelo alcance global da mesma, que possui milhões de usuários cadastrados. Ao contrário de Gomes (2013), que restringe o alcance do estudo a apenas a um lugar. Os textos a serem minerados compõem um *tweet*, que é uma sequência de caracteres publicada pelos usuários, podendo conter outros tipos de dados anexados.

Este trabalho assemelha-se ao de Barbosa et al (2012), por utilizar *hashtags* para determinar o sentimento expresso em um *tweet*, correlacionando os resultados aos fatos ocorridos no período de Copa. Entretanto, difere-se por considerar outras palavras do *tweet* que possam expressar um sentimento, mesmo que estas não estejam marcadas como uma *hashtag*.

3.2 Análise de Sentimentos na classificação de textos

Gomes (2013) ressalta que, com a popularização da Internet, as pessoas geram um imenso volume de dados a cada segundo. O desafio é saber como manipular essa grande quantidade de informação gerada e investigar como as organizações podem se beneficiar

dessas informações, considerando que 80% das informações das organizações estão contidas em documentos de texto (TAN, 1999). Em seu trabalho, Gomes (2013) analisa títulos de notícias sobre economia extraídas de endereços *feeds RSS*. O autor propõe um modelo de Análise de Sentimentos que polariza as notícias em positivas, negativas ou neutras, além de fornecer um documento que guie as organizações a como procederem para extrair conhecimento de dados textuais.

Neste trabalho, as opiniões dos usuários do Twitter sobre a Copa do Mundo serão polarizadas utilizando a mesma categorização: positivas, negativas, neutras ou ambíguas. Entretanto, a classificação será feita por meio de uma implementação do algoritmo de classificação Naive Bayes, que será explicado na subseção 5.3.

3.3 Twitter

Rodrigues Barbosa et al (2012) pontuam que o modelo de interação do Twitter induz os usuários a compartilhar e expressar continuamente suas opiniões e sentimentos, que são propagados para seus respectivos seguidores. Porém, determinar o sentimento que cada *tweet* expressa pode ser uma tarefa trabalhosa, sujeita a erros e ambiguidades. Para contornar esses desafios, Rodrigues Barbosa et al (2012) exploram *hashtags*. As *hashtags* são utilizadas para determinar o sentimento expresso pelos usuários do Twitter nos *tweets* referentes às eleições presidenciais do Brasil, em 2010. A classificação das *hashtags* em algum sentimento que indique sua polaridade é realizada de maneira manual.

Russel (2013) cita a curiosidade humana e a necessidade de compartilhar ideias e experiências, fazer perguntas, interagir de maneira rápida. O Twitter propicia de maneira dinâmica que todos esses aspectos sejam possíveis de serem realizados na velocidade do pensamento. Além disso, a rede social possui um diferencial das demais redes sociais, pelo seu modelo assimétrico de seguidores, onde qualquer usuário pode ficar por dentro dos últimos acontecimentos, mesmo que ele não siga o autor da postagem, enquanto em outras redes sociais, como o Facebook e LinkedIn, é preciso uma aceitação de conexão entre os seus usuários (RUSSEL, 2013).

Tendo em vista a grande vantagem de utilizar o Twitter como rede social para mineração, este trabalho busca detectar as opiniões dos usuários do Twitter, analisando os *tweets* que falam sobre a Copa do Mundo 2014. Serão analisadas *hashtags* que indiquem os sentimentos dos autores das postagens, além de utilizar de outros atributos de um *tweet* como localização dos *tweets* (latitude e longitude de onde foi postado) para desenvolver mapas de

calor, que indicam onde e quando estão comentando mais sobre a Copa do Mundo de 2014 nos dias de jogo da seleção brasileira.

4 OBJETIVOS

4.1 Objetivo geral

O objeto deste trabalho é identificar o sentimento da população sobre a Copa do Mundo através de suas postagens no Twitter, correlacionando as sensações identificadas com os acontecimentos reais que acontecem juntamente com a Copa do Mundo.

4.2 Objetivos específicos

- Coleta dos dados textuais (*tweets*) do Twitter referentes à Copa do Mundo 2014;
- Pré-processamento dos dados coletados;
- Aplicar a Análise de Sentimentos sobre os *tweets* utilizando uma implementação do algoritmo de classificação Naive Bayes para gerar um modelo de classificação de *tweets*;
- Validar os resultados obtidos na fase de análise via conjunto de teste do passo anterior. Além disso, verificar a sensação prevista na classificação dos *tweets* de acordo com os eventos que estão acontecendo ao longo da Copa.
- Mostrar como utilizar o Twitter para analisar as opiniões da população baseado em suas postagens na rede social;
- Gerar mapas de calor, a partir das coordenadas coletadas, que indicam lugares onde mais se falou bem ou mal da Copa nos dias de jogo da Seleção Brasileira de Futebol (território brasileiro);
- Gerar nuvens de palavras, que mostram quais foram as palavras mais faladas nos dias de jogo da Seleção Brasileira de Futebol.

5 PROCEDIMENTOS METODOLÓGICOS

5.1 Coleta de *tweets*

A coleta dos dados do Twitter foi a primeira parte da execução deste trabalho. Para esta etapa, foi utilizado um *script* escrito na linguagem Python. O script recebe como parâmetro *hashtags*, e retorna *tweets* marcados com essas *hashtags*. Esses *tweets* foram em seguida armazenados em arquivos com valores separados por tabulação. Este processo de coleta foi iniciado no início dos jogos (12 de junho) e encerrado no último dia de jogos (13 de julho), para que assim, fosse feita uma análise dos acontecimentos relacionados ao evento durante todo o seu período de acontecimento.

Figura 4 - Atributos de um tweet

Fonte: elaborado pelo autor.

A Figura 3 apresenta um exemplo de um *tweet*, com parte das informações que ele contém: id, texto, número de vezes que foi curtido, coordenadas, dentre outras. Para este trabalho, foram armazenados apenas os atributos que foram considerados relevantes para a execução do mesmo. São eles: id do *tweet*, texto do *tweet*, coordenadas (serão utilizados para gerar mapas de calor) e data de criação do *tweet*.

Ao final desta etapa, iniciou-se a etapa de pré-processamento, que é explicada na subseção a seguir.

5.2 Pré-processamento dos *tweets* coletados

Nesta etapa, os *tweets* passaram por um pré-processamento, onde foram descartados conteúdos considerados irrelevantes para o processo de classificação de texto, como *links*, nomes de usuário (no Twitter, marcados pelo caractere ‘@’) e caracteres não alfabéticos, com exceção do caractere ‘#’, que é utilizado para marcar uma palavra como *hashtag*.

Ainda nesta fase foi aplicada a remoção de *stopwords*, uma técnica de Processamento de Linguagem Natural para a retirada de *stopwords*. Assim, palavras que não possuem relevância para os resultados da classificação de textos foram descartadas dos conjuntos de dados, melhorando o desempenho do algoritmo de classificação de textos que será utilizado. Para a remoção dos stopwords, foi utilizada a plataforma NLTK (*Natural Language ToolKit*).

Ela é uma plataforma para construção de programas em Python que trabalham com dados da linguagem humana[7].

Depois de realizada esta estruturação nos dados coletados, seguimos para a classificação dos dados, utilizando a Análise de Sentimentos. Esta etapa é detalhada na próxima subseção.

5.3 Mineração de Textos

Nesta etapa, o algoritmo de classificação de textos Naive Bayes do Apache Mahout foi utilizado para gerar o modelo de classificação de *tweets* da Copa.

Para isso, foi separada uma pequena quantidade dos *tweets* coletados que foi utilizada para gerar o modelo de classificação. Foi realizada uma classificação manual neste conjunto. Em seguida, dividiu-se o mesmo em dois outros conjutos, denominados treino e teste. A partir do conjunto de treino, foi criado um modelo de classificação dos *tweets*, que associa as palavras às categorias que elas acontecem neste conjunto de treino, para que assim o algoritmo aprenda a rastrear as possibilidades de classificação dos *tweets* em um sentimento, dado a presença das *hashtags* e outras palavras associadas a esse *tweet*. Em seguida, foi validado o modelo de classificação gerado com o conjunto de treino utilizando o conjunto de teste. Assim pode-se verificar quão bom é o modelo obtido. Para complementar esta validação, uma amostra de *tweets* foi classificada, e conferida pelo autor deste trabalho, a fim de coincidir o resultado da acurácia da classificação dos *tweets* da amostra com o resultado da acurácia da classificação do conjunto de teste. O modelo analisa os *tweets* com base nas palavras e *hashtags* que os compõem, podendo ser classificados como positivos ou negativos. Além disso, os *tweets* puderam ser classificados como ambíguos, quando eles podem ser associados a algum sentimento, mas não de maneira clara, ou neutros, quando não for possível associá-los a algum sentimento.

[7] http://www.nltk.org/

Finalizado este processo, partiu-se para a etapa de avaliação do resultado da classificação em relação aos eventos que ocorrem juntamente com a Copa, detalhado na próxima subseção.

5.4 Análise de Sentimentos

5.4.1 Classificação

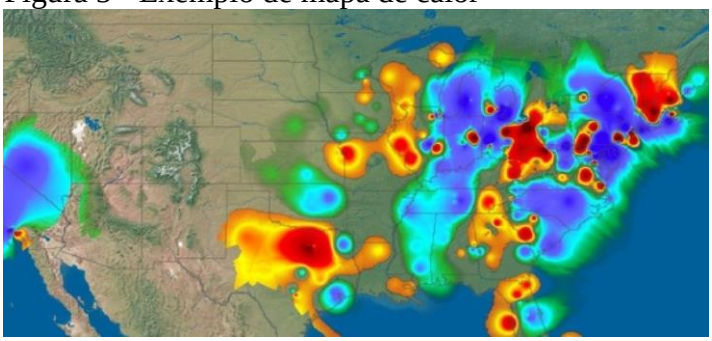
Após a validação da classificação do modelo de treino gerado na etapa anterior, foi realizada a classificação dos demais *tweets*, que foram categorizados de acordo com uma das classes definidas para o nosso modelo de classificação: positivo, negativo, ambíguo ou neutro. Após a classificação, foi realizada uma análise dos fatos que ocorreram no período da Copa em relação com a classificação realizada através do modelo de classificação. A classificação foi então validada com base nos fatos que ocorreram correlacionados à Copa do Mundo 2014, como resultados de partidas, manifestação das torcidas a respeito dos jogos, manifestações contra a Copa, dentre outros fatos.

5.4.2 Geração de mapas de calor

Os mapas de calor podem ser descritos como representações gráficas de dados. Os mapas de calor gerados neste trabalho mostram em que regiões do Brasil mais se falaram bem e mal da Copa do Mundo nos dias de jogos em que a seleção brasileira jogou, baseados na classificação dos *tweets* coletados que possuíam informação de coordenadas, utilizadas para gerar os mapas. Para a geração dos mapas de calor, foi utilizada a API Javascript do Google Maps V3. A Figura 5 ilustra um mapa de calor. A API oferece diversos utilitários para a manipulação de mapas e para a adição de conteúdo ao mapa por meio de diversos serviços[8].

[8] https://developers.google.com/maps/documentation/javascript/?hl=pt-br

Figura 5 - Exemplo de mapa de calor



Fonte: Silicon Graphics International – Repositório Digital do The Huffington Post[9]

5.4.3 Nuvens de palavras

Por fim, este trabalho apresenta nuvens de palavras, imagens compostas por palavras, para demonstrar de maneira visual, a frequência da ocorrência das palavras em um dado texto: quanto maior for o número de ocorrências de uma palavra, maior a mesma será na nuvem de palavras. Este resultado apresenta como resultado a análise da alternância das *hashtags* e demais palavras mais frequentes ao longo dos dias das partidas. Um exemplo de nuvem de palavras é apresentado na Figura 6 (abaixo). Na ocasião, a nuvem foi gerada a partir dos textos coletados no dia de estreia do evento.

Figura 6 - Exemplo de nuvem de palavras

Fonte: elaborado pelo autor.

[9] http://www.huffingtonpost.com/2012/11/19/sgi-twitter-heat-map_n_2138726.html. Acesso em abr. 2014.

6 RESULTADOS

6.1 Coleta de dados

Durante o período de coleta de *tweets*, foram coletados 2.128.862 *tweets* distintos. Para isso, foram observados *hashtags* e *trending topics* correlacionados à Copa do Mundo e aos jogos, ao longo período definido, para que fosse possível coletar os *tweets* que falavam sobre o evento. Os *tweets* coletados foram salvos em planilhas com valores separados por tabulação. Os *tweets* foram separados em planilhas diferentes, de acordo com cada dia de coleta.

Tabela 1 - Informações sobre etapa de coleta de tweets

Descrição	Quantidade
Número de dias	32
Tweets coletados	2.128.862
Tweets coletados (com geolocalização)	61.000
Tweets (conjunto de treino)	3.250
Tweets (amostra)	312
Tamanho total em disco (arquivos TSV)	868 *MegaBytes*
Tamanho total em disco (arquivos TSV com *tweets* pré-processados)	276 *MegaBytes*

Fonte: elaborado pelo autor.

6.2 Pré-processamento dos dados coletados

Após a fase de coleta, foram removidos de todos os textos dos *tweets* coletados conteúdos não relevantes para a fase de mineração: links, nomes de usuário do Twitter, *stopwords*, caracteres especiais e caracteres numéricos, salvo o caractere '#', por marcar uma palavra como hashtag.

6.3 Desenvolvimento e validação do modelo de classificação

Após o pré-processamento dos textos coletados, foram selecionados de todas as planilhas, de maneira randômica, 3285 *tweets*, para gerar o modelo de classificação. Os *tweets* desse conjunto foram classificados manualmente pelo autor deste trabalho, de acordo com uma das quatro polaridades definidadas (positivo, negativo ambíguo e neutro). Em seguida,

esse conjunto de treino classificado foi dividido em duas partes, onde 80% dos *tweets* foram selecionados randomicamente para treinar o algoritmo de Naive Bayes (2634 *tweets*) e 20% dos *tweets* (651 *tweets*) foram selecionados randomicamente para testar o modelo.

Após gerado o modelo, foi testado a acurácia de classificação do mesmo. O modelo apresentou uma taxa 88,91% de precisão, utilizando o conjunto de treino contendo 2634 *tweets* para testá-lo. O mesmo apresentou ótimas taxas de precisão para a categoria positiva (94%) de precisão, negativa (93%) e neutra (84%). Para classe ambígua, mostrou-se bom (75%).

Em contrapartida, ao ser testado com o conjunto de teste (651 *tweets*), o modelo apresentou uma taxa de 74,80% de precisão. O modelo apresentou 82% de precisão para a categoria positiva, 84% de precisão para a categoria negativa, 69% para a categoria neutra e apenas 40% de precisão para a categoria ambígua.

O modelo, assim, apresentou-se bom para classificar *tweets* positivos e negativos. No ambíguo e neutro, ocorreram uma maior taxa de erro.

Foi observado um grande número de ocorrências em que um *tweet* continha mais de uma *hashtag*, onde cada uma expressava um sentimento diferente (às vezes oposto), dificultando o trabalho de classificação para o conjunto de treino, sendo esses, uma das dificuldades encontradas para se gerar um melhor modelo de classificação.

Também fora submetida uma pequena amostra com 312 *tweets* para serem classificadas com o modelo gerado. Deste número, 167 instâncias são positivas. Foi observado que o modelo classificou corretamente 98% desta quantia (165 instâncias), mostrando-se muito eficiente para *tweets* positivos. Já, na categoria negativa, de 85 instâncias, o modelo classificou corretamente 70 (76% do total de negativas). Para a classe ambígua, de 22 instâncias, o modelo classificou corretamente 8 delas (36% do total de ambíguas), sendo esta a categoria onde a classificação foi muito ruim. Já na neutra, o modelo classificou corretamente 24 de 32 instâncias (75% do total de neutras). No geral, o modelo gerado classificou corretamente 85,57% das instâncias que compunham a amostra.

6.4 Análise de sentimentos

Realizada a validação do modelo, partimos em seguida para a classificação dos demais *tweets* utilizando nosso modelo de classificação, bem como a geração das nuvens de palavras e mapas de calor dos dias de jogos da seleção brasileira. Foi observado pelo autor

deste trabalho, que a quantidade de *tweets* nos dias de jogos da seleção brasileira era bem superior a quantidade de *tweets* coletdos nas demais datas. Assim, A Tabela 2 (abaixo) foi criada para ilustrar os dias de partida da seleção brasileira de futebol e resultados dos jogos, para melhor acompanhar os comentários nas análises seguintes.

Tabela 2 - Datas e resultados das partidas da Seleção Brasileira de Futebol

Partida	Data	Resultado
Brasil x Croácia (1º fase)	12/06/2014	3 x 1
Brasil x México (1º fase)	17/06/2014	0 x 0
Camarões x Brasil (1º fase)	23/06/2014	1 x 4
Brasil x Chile (Oitavas de Final)	28/06/2014	1 x 1 (3 x 2 pênaltis)
Brasil x Colômbia (Quartas de Final)	04/07/2014	2 x 1
Brasil x Alemanha (Semifinal)	08/07/2014	1 x 7
Brasil x Holanda (3º Lugar)	12/07/2014	0 x 3

Fonte: elaborado pelo autor.

6.4.1 Classificação

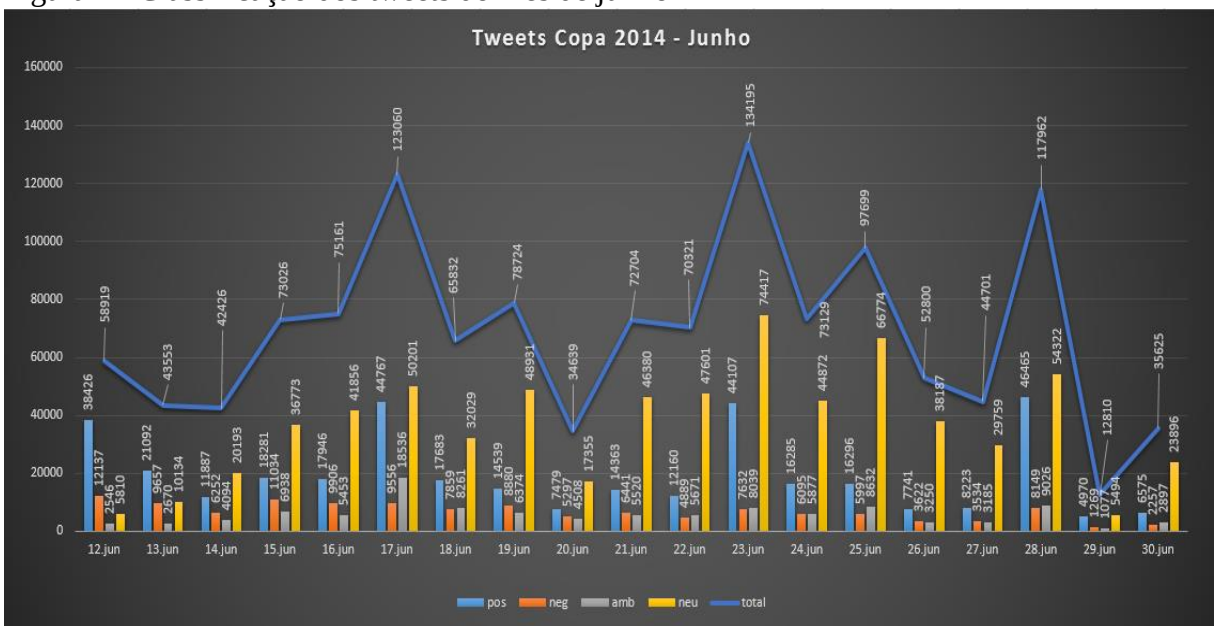
A partir do modelo de classificação *tweets* criado, classificamos todos os 2.128.862 *tweets* coletados e pré-processados.

A Figura 7 (próxima página) ilustra o resultado da classificação dos *tweets* no mês de junho. Podemos observar que a quantidade de *tweets* classificados como positivos foi muito superior à quantidade de *tweets* classificados como negativos, sendo bastante expressiva no dia 12.06 (estréia), caindo um pouco até os dias 16.06 (um dia antes da partida da seleção brasileira) e 17.06 (dia de jogo da seleção brasileira), onde voltaram a ser expressivos, mostrando assim boas expectativas dos usuários do Twitter próximos aos dias de jogos da seleção brasileira. Na ocasião, a seleção brasileira jogou no dia 17.06 contra a seleção mexicana, empatando a partida em 0x0[10]. A situação se repetiu nos outros dias em que a seleção brasileira joga. Também podemos conferir a mesma situação na Figura 8 (mais abaixo), que ilustra o resultado da classificação dos *tweets* no mês de julho.

[10] http://globoesporte.globo.com/futebol/copa-do-mundo/noticia/2014/06/ochoa-brilha-brasil-empata-sem-gols-com-o-mexico-mas-ainda-lidera.html

Como se pode observar nas Figuras 7 e 8, a quantidade de *tweets* ambíguos foi relativamente inexpressiva, enquanto houve um grande número de *tweets* classificados como neutros. Uma grande quantidade de *tweets* que noticiavam partidas, e que possuíam muito *retweets* foi um dos motivos para esta classe ter um número expressivo de *tweets* classificados como neutros. Temos como exemplo de tweet classificado como neutro e muito "retuitado", o *tweet* "conheca novo aplicativo band nao perca nenhum lance copa".

Figura 7 - Classificação dos tweets do mês de junho



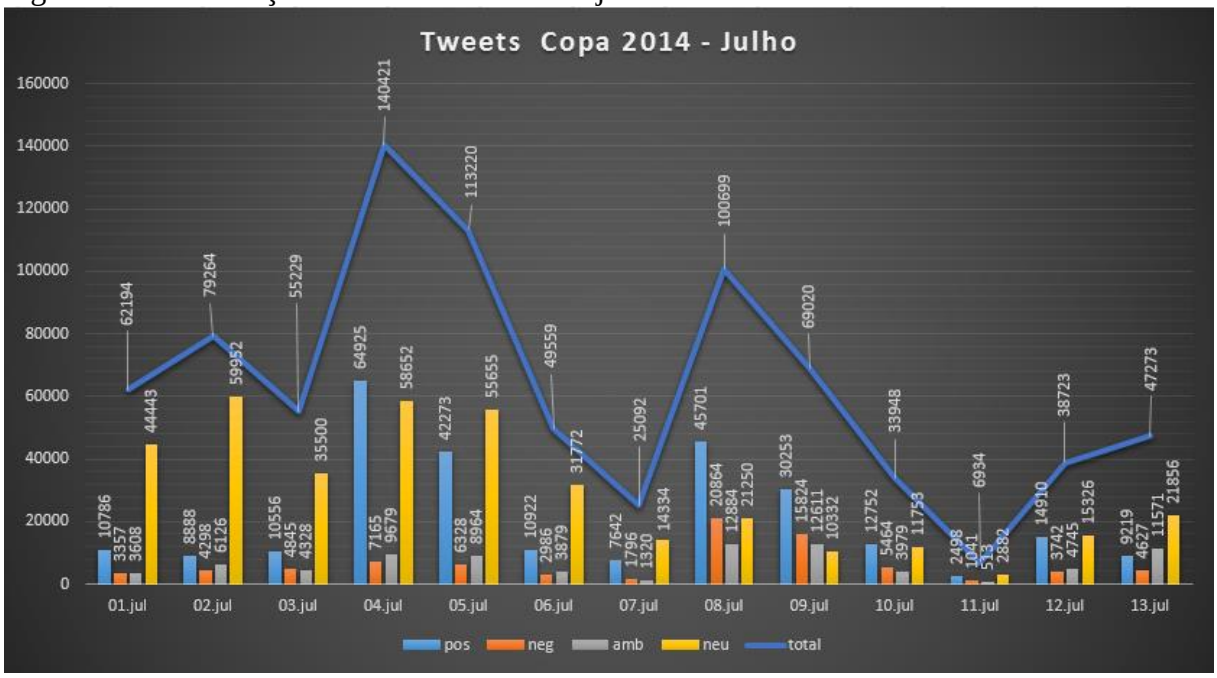
Fonte: elaborado pelo autor.

A quantidade de *tweets* positivos mostrou-se expressiva ao longo de todos os dias da Copa, especialmente próximo das partidas e nos dias de partida da seleção brasileira, mostrando assim que os usuários do Twitter foram favoráveis à Copa. Esse comportamento favorável dos usuários do Twitter se refletiu no país. O número de manifestações caiu após o início do evento[11][12], não repetindo a onda de manifestações ocorridas em junho de 2013, quando o país sediou a Copa das Confederações da FIFA, outro torneio futebolístico organizado pela FIFA. A imprensa internacional elogia a hospitalidade do povo brasileiro e o bom comportamento nos estádios[13].

[11] http://agenciabrasil.ebc.com.br/geral/noticia/2014-06/manifestacoes-diminuiram-na-copa-do-mundo
[12] http://www1.folha.uol.com.br/poder/2014/06/1475841-numero-de-manifestacoes-cai-39-apos-o-inicio-da-copa-do-mundo.shtml
[13] http://globoesporte.globo.com/futebol/copa-do-mundo/noticia/2014/07/apos-espera-de-fiasco-imprensa-muda-discurso-e-copa-e-sucesso-fora-do-pais.html

Podemos observar na Figura 8 (abaixo), os *tweets* coletados no período final do evento. O número de *tweets* classificados como negativos sobe consideravelmente em relação aos dias de partida anteriores. Na ocasião, a seleção brasileira perdeu a sua primeira partida já na semifinal do evento, contra a seleção alemã por 7x1, maior derrota já registrada na história do time[14]. Entretanto, a quantidade de *tweets* positivos, mesmo diante a derrota por goleada é grande. Fato motivado pelo número de *tweets* com palavras de apoio ao jogador da seleção brasileira David Luiz, considerado destaque do time no evento, ganharam os assuntos do momento no Twitter, após a derrota da seleção. Alguns noticiários o trataram como o jogador mais querido do Brasil[15].

Figura 8 - Classificação dos tweets do mês de julho



Fonte: elaborado pelo autor.

6.4.2 Mapas de calor

No intuito de mapear as opiniões por localidade, foram mantidos os atributos de localização, contidos nos *tweets*, para que fossem gerados mapas de calor. Para que tal informação seja adicionada ao *tweet*, entretanto, o usuário deve ativar o recurso de localização do Twitter no momento em que realizará a postagem, para que o mesmo possa inserir informação de localização da publicação do *tweet.* Sendo assim, apenas os *tweets* em que os

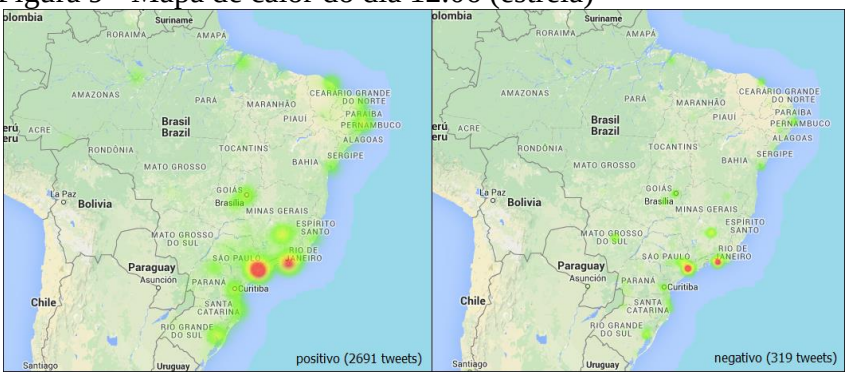
[14] http://placar.abril.com.br/materia/derrota-para-a-alemanha-foi-a-10a-maior-goleada-em-copas
[15] http://trivela.uol.com.br/david-luiz-o-jogador-mais-amado-brasil-em-6-imagens/

usuários ativaram este recurso vieram com informação de localização, correspondendo a apenas um pequeno número dos *tweets* coletados (quantidade descrita na subseção 6.1).

Podemos observar ao longo das Figuras 9, 10 e 11 (abaixo) que as postagens do Twitter vinham dos grandes centros do Brasil, principalmente de estados que receberam partidas.

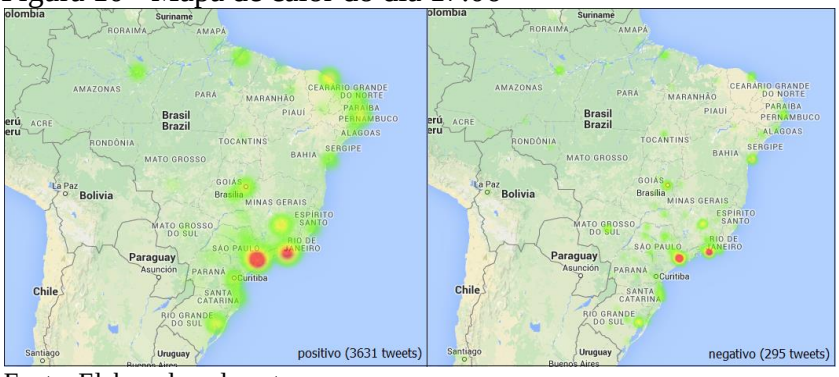
Em todos os dias em que aconteceram partidas da seleção brasileira, a quantidade de *tweets* positivos foram superiores ao número de *tweets* negativos, que mais uma vez se mostram inexpressivo. Os pontos mais fracos (verdes) indicam menor ocorrência de *tweets* sobre a região que se encontram.

Figura 9 - Mapa de calor do dia 12.06 (estréia)



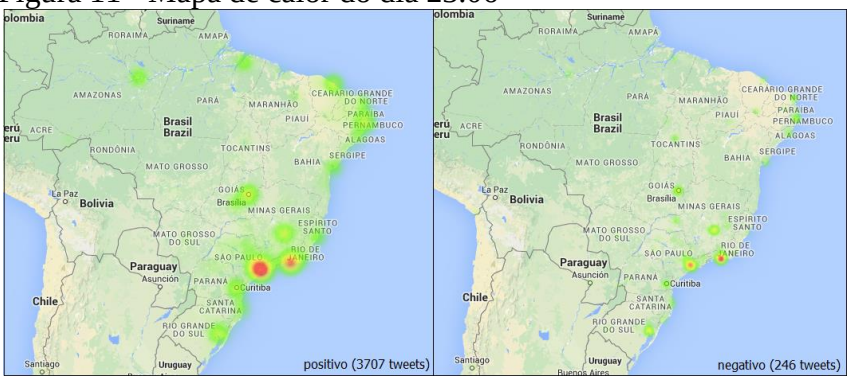
Fonte: elaborado pelo autor.

Figura 10 - Mapa de calor do dia 17.06



Fonte: Elaborado pelo autor

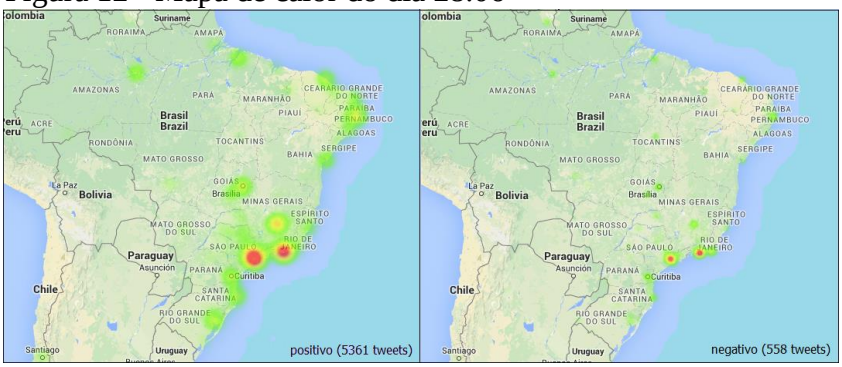
Figura 11 - Mapa de calor do dia 23.06



Fonte: elaborado pelo autor.

As regiões Sudeste e Sul foram as que mais postaram *tweets* de ambos os teores (positivo e negativo), e as postagens se intensificaram nas últimas partidas. Os pontos mais fortes nos mapas (vermelhos) indicam maior ocorrência de *tweets* sobre a região que se encontram.

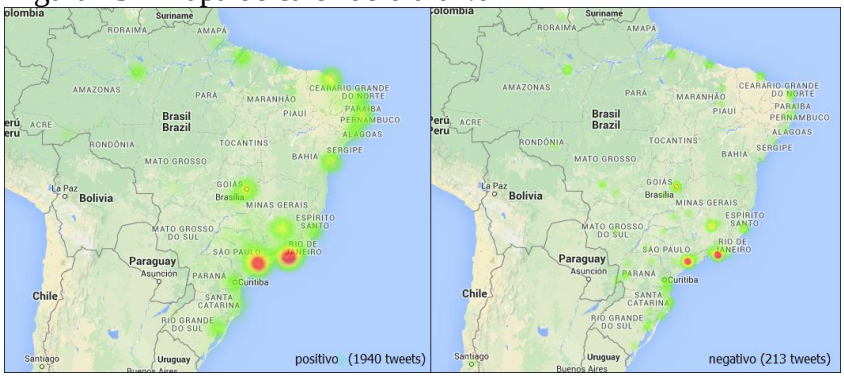
Figura 12 - Mapa de calor do dia 28.06



Fonte: elaborado pelo autor.

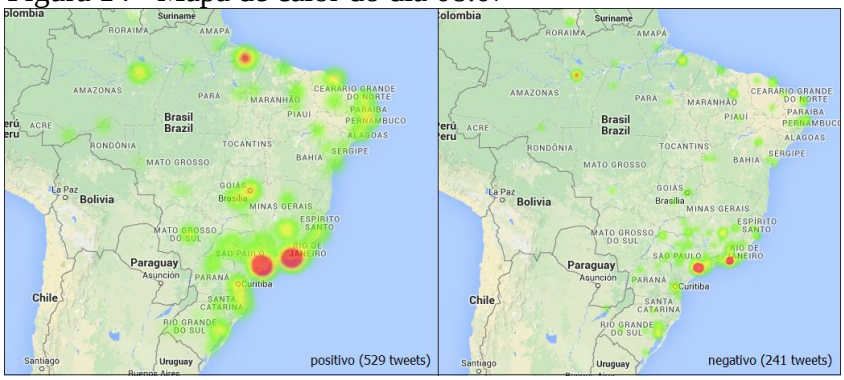
Podemos observar que, a partir do dia 04.07 (Figura 13), a quantidade de *tweets* com localização foi reduzindo, à medida que os números de partidas foram diminuindo, já que esse foi o período de quartas de final as finais dos jogos.

Figura 13 - Mapa de calor do dia 04.07



Fonte: elaborado pelo autor.

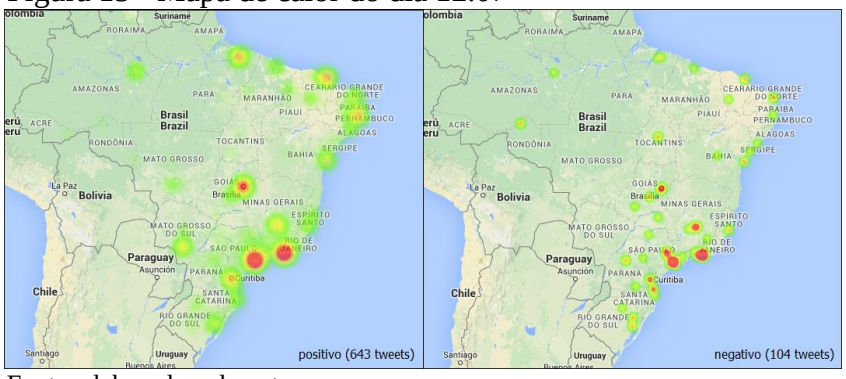
Figura 14 - Mapa de calor do dia 08.07



Fonte: elaborado pelo autor.

Os lugares com maior número de postagens positivas e negativas nos dias de jogos da seleção brasileira são pertencentes a região sudeste do país, como São Paulo e Rio de Janeiro, estados que possuíam cidades-sede do evento. Pode-se assim afirmar que esta foi a região onde se teve maior quantidade de usuários do Twitter falando sobre a Copa do Mundo durante os jogos da seleção brasileira.

Figura 15 - Mapa de calor do dia 12.07

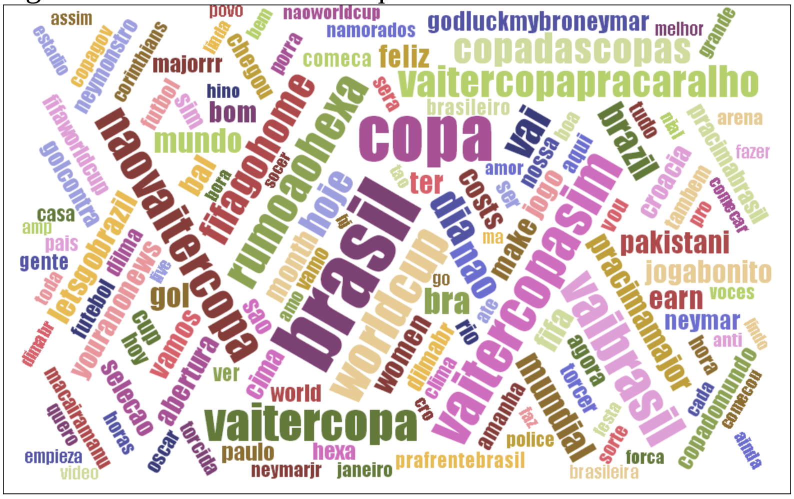


Fonte: elaborado pelo autor.

6.4.3 Nuvens de palavras

Como mencionado, foram geradas nuvens de palavras para cada dia de jogo, no intuito de representar graficamente as variações das palavras mencionadas com mais frequência ao longo dos dias de partidas. É possível observar na Figura 16 (abaixo), dia de estreia da seleção brasileira na Copa do Mundo, uma grande alternância de palavras que expressam um sentimento positivo, como “vaibrasil” e ‘rumoaohexa’, com palavras negativas, fazendo alusões contra a realização do evento, como “naovaitercopa” e “fifagohome”.

Figura 16 - Palavras mais frequentes do dia 12/06/2014 –BRA x CRO



Fonte: elaborado pelo autor.

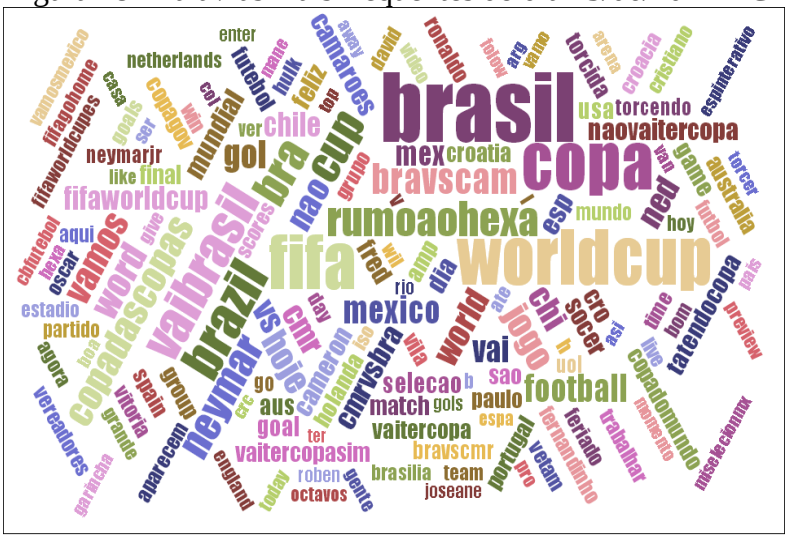
A seleção brasileira venceu a seleção croata na estreia. Entretanto, o primeiro gol da seleção brasileira foi contra, gerando um furor entre os usuários no Twitter. Podemos observar, a palavra “golcontra” aparece entre as mais frequentes.

As palavras de apoio seguem até o último dia de partida da seleção nacional, como é possível observar nas Figuras 17, 18, 19, 20, 21 e 22, enquanto as que fazem alusão contra a Copa vão se destacando menos.

Figura 17 - Palavras mais frequentes do dia 17/06/2014 –BRAx MEX

Fonte: elaborado pelo autor.

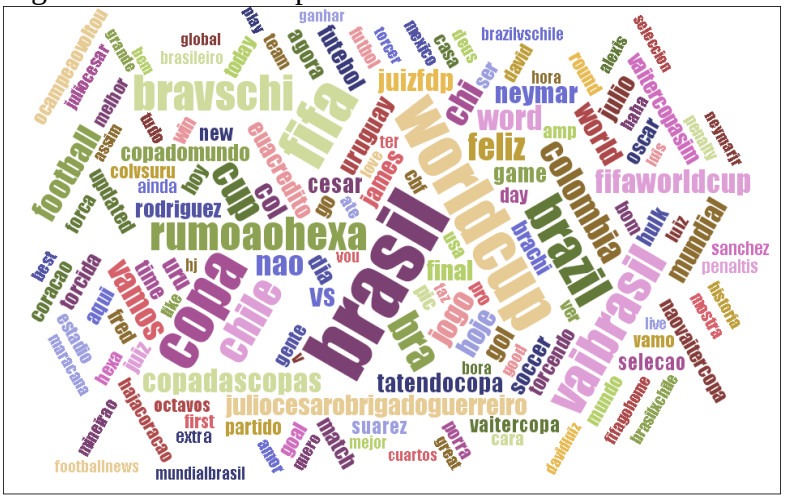
Figura 18 - Palavras mais frequentes do dia 23/06/2014 – CAM x BRA



Fonte: elaborado pelo autor.

Pode-se observar que em todas as nuvens de palavra aparecem menções ao jogador Neymar Júnior, como nas Figuras 18 (acima), 19 e 20 (abaixo). O mesmo é considerado um dos melhores jogadores da seleção brasileira. No dia do jogo contra a seleção colombiana, o mesmo sofreu uma lesão após uma falta grave[16]. Pode-se observar as menções "neymarjr" e "neymar" entre os assuntos de maior destaque na nuvem de palavras referentes ao dia daquele jogo (mais abaixo, na Figura 20), mostrando a grande popularidade do mesmo.

Figura 19 - Nuvem de palavras do dia 28/06/2014 – BRA x CHI



Fonte: Elaborado pelo autor

[16] http://copadomundo.uol.com.br/noticias/redacao/2014/07/04/neymar-esta-fora-da-copa-do-mundo-aposfraturar-vertebra.htm

Figura 20 - Nuvem de palavras do dia 04/07/2014 – BRA x COL

Fonte: elaborado pelo autor.

No dia 08.07, palavras com teor negativo como “vergonha” e “vergonhabrasil” ganham destaque (Figura 24). O número de *tweets* classificados como negativos sobe consideravelmente em relação aos dias de partida anteriores, como se pode conferir na Tabela 7. Na ocasião, a seleção brasileira perdeu a sua primeira partida já na semifinal do evento, contra a seleção alemã por 7x1, maior derrota da história da seleção brasileira. Em mesma proporção, palavras de apoio ao jogador da seleção brasileira David Luiz, considerado destaque do time no evento, ganharam os assuntos do momento no Twitter, após a derrota da seleção.

Houve rumores de que a Copa haveria sido comprada e árbitros das partidas da seleção brasileira favoreceriam a equipe em seus jogos[17]. Com a derrota, usuários do Twitter brincaram com o assunto, mencionando em seus *tweets* a *hashtag* “#dilmaesqueceudepagaralemanha”, culpando a presidente do Brasil Dilma Roussef pela derrota do time (Figura 21).

[17] http://copadomundo.uol.com.br/noticias/redacao/2014/06/29/jornal-alemao-afirma-que-jogo-do-grupo-do-brasil-pode-ter-sido-comprado.htm

Figura 21 - Palavras mais frequentes do dia 08.07 – BRA x ALE

Fonte: elaborado pelo autor.

No último dia de partida da seleção brasileira, disputando o 3º lugar do mundial contra a seleção holandesa, palavras de apoio ao jogador David Luiz continuam em alta no Twitter. Palavras negativas como ‘vergonha’ continuaram entre as mais mencionadas. A seleção, neste dia, perdeu novamente, ficando com o 4º lugar da Copa do Mundo[18]. Observe que na Figura 22, palavras como ‘presosdacopa’ e ‘liberdade’ ficaram em alta no último dia de participação da seleção brasileira na Copa. Nesse mesmo dia, policiais civis da cidade do Rio de Janeiro realizaram a prisão de 19 manifestantes, suspeitos de atos de vandalismos durante manifestações desde junho de 2013[19], ano em que ocorreu a Copa das Confederações, evento futebolístico promovido pela FIFA, no país. Vários usuários do Twitter manifestaram-se contra as prisões dos suspeitos presos.

[18] http://www.jj.com.br/noticias-3091-vaiada-selecao-brasileira-perde-para-a-holanda-por-3-a-0
[19] http://noticias.uol.com.br/cotidiano/ultimas-noticias/2014/07/12/policia-civil-do-rj-prende-19-manifestantes-um-dia-antes-da-final-da-copa.htm

Figura 22 - Nuvem de palavras do dia 12/07/2014 – BRA x HOL

Fonte: elaborado pelo autor.

Foi possível perceber como o conteúdo das redes sociais refletiu o sentimento e o pensamento da população brasileira durante o evento.

7 TRABALHOS FUTUROS

Foi utilizado neste trabalho o algoritmo de classificação de textos Naive Bayes. O mesmo é de fácil implementação e muito eficiente.

Na busca de obter melhores resultados, é possível aperfeiçoar o modelo de classificação, retroalimentando o conjunto de treino com mais *tweets*, balanceando o número de *tweets* para cada classe. Outras categorias também podem ser definidas, de acordo com o contexto dos dados, utilizando o algoritmo Naive Bayes.

Além disso, outras técnicas de classificação podem ser utilizadas e comparadas, como classificadores de textos utilizando Máquinas Vetores de Suporte (do inglês *Support Vectos Machines*). As Máquinas de Vetores de Suporte constituem uma técnica de aprendizado de máquinas com base na teoria de aprendizado estatístico, desenvolvida por Vapnik (1995). Essa teoria, por sua vez, estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização (LORENA, 2006).

Foi empregada apenas uma técnica de Processamento de Linguagem Natural utilizando a ferramenta NLTK. Existem outras técnicas que podem ser empregadas para melhor estruturar os textos que serão minerados, no intuito de se obter melhores resultados, como Reconhecimento de Entidades Nomeadas, técnica que consiste em identificar entidades nomeadas, na sua maioria nomes próprios, a partir de textos de forma livre e classificá-las dentro de um conjunto de tipos de categorias pré-definidas, tais como pessoa, organização e localização (,). Os textos extraídos do Twitter utilizados neste trabalho podem vir com nome de pessoas, o que não era relevante para os resultados do trabalho, mas poderiam vir a interferir nos mesmos. Assim, pode-se utilizar o Reconhecimento de Entidades Nomeadas para identificar nomes indesejados nos textos e extraí-los, para que não interfiram na etapa de mineração.

Pode-se experimentar o processo seguido neste trabalho em outras redes sociais, como o Facebook, Whatsapp (histórico de conversas), dentre outras. Como exemplo, pode-se utilizar as etapas de mineração de textos desse trabalho para gerar um modelo de classificação que permita extrair conhecimento a partir de comentários em páginas populares do Facebook.

8 CONCLUSÃO

Este trabalho apresentou como o processo de Mineração de Textos foi usado para coletar, estruturar o texto extraído do Twitter (*tweets*) e como criar um modelo de classificação de texto para os *tweets* que falavam sobre a Copa, que permitiu conhecer a opinião da rede social do usuário do Twitter sobre Copa do Mundo da FIFA Brasil 2014. As postagens dos usuários, popularmente conhecido como tweets, foram categorizadas neste trabalho em um sentimento: positivo, negativo, ambíguo ou neutro.

Assim, o modelo de classificação gerado neste trabalho nos permitiu mostrar a opinião dos usuários ao longo do período da Copa, validando as classificações feitas pelo modelo com os fatos associados a Copa no país, como o fato da seleção ter sido desclassificada na semifinal contra a seleção alemã, após perder de goleada, influenciou no aumento do número de *tweets* negativos, que até então eram inexpressivos.

Além disso, foram apresentados neste trabalho mapas de calor dos dias de partida da seleção brasileira, que permitiram conhecer em quais regiões se comentavam mais sobre a Copa, e onde mais falavam bem ou mal do evento. Também foram apresentadas nuvens de palavras, que permitiram saber quais eram as palavras mais citadas pelos usuários do Twitter nos dias de partida da seleção brasileira, correlacionando as palavras mais frequentes com fatos que ocorreram relacionados à Copa.

Este trabalho teve como fruto o artigo Análise de Sentimentos de *tweets* nos dias de jogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014 utilizando Mineração de Textos (FILHO; LEITE; DA SILVA, 2014), aceito e apresentado no evento ENUCOMP 2014[20].

Assim, o processo apresentado neste trabalho, pode ser seguido por organizações para mapear a opinião de usuários do Twitter, fazendo o uso dos resultados para os mais diversos fins dentro das mesmas.

O processo de classificação de textos utilizado neste trabalho está sendo aplicado na empresa iFactory Solutions, para analisar a opinião de usuários sobre os clientes da empresa nas redes sociais.

[20] http://www.enucomp.com.br/2014/artigos

REFERÊNCIAS

ARANHA, C.N. *Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional*. 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2007.

DO AMARAL, D. O. F. *O reconhecimento de entidades nomeadas por meio de conditional Random Fields para a língua portuguesa*. 2013. 99 f. Dissertação (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre. 2013.

EL-KHAIR, Ibrahim Abu. Effects of stop words elimination for Arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences*, v. 4, n. 3, p. 119-133, 2006.

FELDMAN, Ronen et al. Knowledge Management: A Text Mining Approach. In: **PAKM**. 1998.

FILHO, José Adail Carvalho; DA SILVA, Ticiana Linhares Coelho; LEITE, João Lucas Araújo. Análise de Sentimentos de *tweets* nos dias de jogos da Seleção Brasileira de Futebol na Copa do Mundo da FIFA Brasil 2014 utilizando Mineração de Textos. In: **Encontro Unificado da Computação**, VII. Parnaíba. 2014.

GOMES, G. R. R. *Integração de Repositórios de Sistemas de Bibliotecas Digitais e de Sistemas de Aprendizagem.* 2006. 143 f. Tese (Doutorado em Informática) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2006.

GOMES, Helder Joaquim Carvalheira. Text Mining: análise de sentimentos na classificação de notícias. **Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on**. Lisboa. 2013.

HEARST, M. A. Untangling text data mining. **Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics** (pp. 3–10),1999. Stroudsburg, PA, USA: Association for Computational Linguistics.

INDURKHYA, Nitin; DAMERAU, Fred J. *Handbook of natural language processing*. 2ed. Florida: CRC Press, 2010.666 p.

LIDDY, E. *Natural Language Processing*. Encyclopedia of Library and Information Science. New York: Marcel Decker, Inc, 2001

LORENA, Ana Carolina. *Investigação de estratégias para a geração de máquinas de vetores de suporte multiclasses*. 2006. 229 f. Tese (Doutorado em Ciência da Computação e Matemática Computacional) – Universidade de São Paulo, São Paulo. 2006.

MORAIS, Edilson Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de Textos**. Goiânia: UFG. 2007. (Série Texto Técnico, INF_005/07)

OWEN, Sean et al. *Mahout in Action*. Connecticut: Manning Publications Co, 2011. 373p.

RODRIGUES BARBOSA, Glívia Angélica et al. Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In: PROCEEDING OF THE 2012 ACM ANNUAL CONFERENCE EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS EXTENDED ABSTRACTS**.** Austin, 2012.

RENNIE, J. D. et al. Tackling the poor assumptions of naive bayes text classifiers. In: ICML. 2003. p. 616-623.

RUSSEL, Mathew A. *Mining the social web*: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More. 2 ed. Sebastopol: O'reilly Media, Inc., 2013.

TAN, Ah-Hwee. Text mining: The state of the art and the challenges. In: PROCEEDINGS OF THE PAKDD 1999 WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, Beijing. 1999.

VAPNIK, Vladimir. *The nature of statical learning theory*. New York: Springer-Verlag, 1995.