UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM ENGENHARIA DE SOFTWARE

JOÃO LUCAS ARAÚJO LEITE

MINERAÇÃO DE TEXTOS DO TWITTER UTILIZANDO TÉCNICAS DE CLASSIFICAÇÃO

QUIXADÁ
2015

JOÃO LUCAS ARAÚJO LEITE

MINERAÇÃO DE TEXTOS DO TWITTER UTILIZANDO TÉCNICAS DE CLASSIFICAÇÃO

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 Prof[a]. Ticiana Linhares Coelho da Silva

QUIXADÁ
2015

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

L555m Leite, João Lucas Araújo
Mineração de textos do Twitter utilizando técnicas de classificação / João Lucas Araújo Leite. – 2015.
40 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Bacharelado em Engenharia de Software, Quixadá, 2015.
Orientação: Profa. Me. Ticiana Linhares Coelho da Silva
Área de concentração: Computação

1. Redes sociais. 2. Mineração de dados (Computação) 3. Classificação - Computação I. Título.

CDD **303.483**

JOÃO LUCAS ARAÚJO LEITE

MINERAÇÃO DE TEXTOS DO TWITTER UTILIZANDO TÉCNICAS DE CLASSIFICAÇÃO

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/Tecnólogo.

Área de concentração: computação

Aprovado em: _____ / junho / 2015.

BANCA EXAMINADORA

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

_____
Prof. Msc. Regis Pires Magalhães
Universidade Federal do Ceará-UFC

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

A minha família, por toda confiança e fé que foi confiado
a mim para chegar até aqui, e além.

AGRADECIMENTOS

Aos meus pais, Oriel e Francisca, mesmo distantes ficicamente, fizeram o máximo possível para me dar a melhor educação, amor e carinho.

Ao meu irmão, Gervásio (dé), por ser uma pessoa inspiradora, que não só irmão, mas também meu melhor amigo, que sempre me estendeu a mão, muito obrigado por todos os conselhos que você me deu, te amo muito.

Agradeço ao meu tio João Araújo, por todo apoio que me foi dado, e por me ajudar nos momentos mais difíceis.

Agradeço a minha tia Graça, por todo carinho e amor que a senhora tem por mim, por não medir esforços para me ajudar, por sempre abrir as portas quando preciso, saiba que a considero uma segunda mãe.

Ao Sergio, Adail, Jonas, Darlan, Amaro, Marcilio, Richas, Thiago e Ygor, meus irmãos, que proporcionaram o melhor momento da minha vida durante essa formação, a quem sempre posso contar.

Aos meus amigos Richas, Jéssica, Hinessa, Natália, Marcelo, Uil, Cris, Guylherme, Matheus, Jhon, João, Rafael, Laísa, Isaac, Dyana, Dada que fizeram esses 4 anos os melhores da minha vida.

Agradeço especialmente a minha orientadora Ticiana Linhares, pelo conhecimento e oportunidades que adquiri, 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.

A todos que fizeram parte da minha formação, muito obrigado a todos.

Amo a todos. Amém.

"Amadurecer talvez seja descobrir que sofrer algumas perdas é inevitável, mas que não precisamos nos agarrar à dor para justificas nossa existência."
(Martha Medeiros)

RESUMO

Nos últimos anos, em consequência do crescimento das redes sociais, as pessoas interagem e compartilham informações relevantes na Internet. Esse crescimento proporcionou algumas vantagens. Uma das vantagens é o compartilhamento de informações entre os usuários em um curto espaço de tempo. Alguns usuários das redes sociais costumam postar opiniões sobre diferentes eventos que estão em alta no momento. Consequentemente, o volume diário de dados que é gerado através das redes sociais cresce exponencialmente e toda a informação gerada através destes poderá ser relevante, se for tratada e utilizada corretamente. Desta forma, surge o estímulo de gerar conhecimento a partir destas informações, de forma organizada e automatizada. Assim, o objetivo principal deste trabalho consiste na construção de um modelo capaz de classificar, em categorias, os dados que foram capturados da rede social Twitter. As postagens dos usuários, conhecida como *tweets*, são categorizadas neste trabalho como: economia, esporte, religião, política e outros. Os dados coletados do Twitter passaram por um processamento de linguagem natural antes de ser gerado o modelo, a fim de retirar *stopwords*, acentos, pontuação e sufixos. O modelo obtido apresentou acurácia satisfatória e foi gerado utilizando a técnica de classificação Naive Bayes.

Palavras chave: Classificação. Mineração de texto. Redes sociais.

ABSTRACT

In recent years, due to the growth of social networks, people interact and share relevant information on the Internet. This growth has some advantages. One of the advantages is the sharing of information between users in a short time. Some social networks users often post reviews of different events that are the breaking news in that moment. Consequently, the daily volume of data that is generated through social networks grows exponentially and all the information generated through the may be relevant if it is treated and used correctly. Therefore, there is the stimulus to generate knowledge from this information in an organized and automated. The main purpose of this work is to construct a model to classify into categories, the data that was captured from the social network Twitter. Posts of users, known as tweets, are categorized in this work as: economy, sport, religion, politics and others. The data collected from Twitter went through a natural language processing before building the model, in order to remove stopwords, accents, punctuation and suffixes. The obtained model showed satisfactory accuracy and was generated using the Naive Bayes classification algorithm.

Keywords: Classification. Text mining. Social networks.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas ultimas décadas, o aumento contínuo computacional tem gerado um grande aumento no fluxo de dados (QU et al. 2012). De acordo com Wu et al. (2013), a cada dia 2.5 quintilhões de bytes de dados são criados e 90% dos dados no mundo hoje foram produzidos nos últimos dois anos. Como exemplo dessa enorme geração de dados, a Figura 1 demonstra o efeito do que acontece na Internet em um período de um minuto.

As redes sociais tornaram a comunicação baseada em computador algo bastante simples e rápido de se realizar, permitindo que seus usuários divulguem e compartilhem informações sobre atividades, opiniões e status. Dentre essas redes sociais, o Twitter está entre uma das mais populares. De acordo com Figueiredo e Garcia (2011), o Twitter permite que o usuário crie conteúdos através de publicações, com um limite de 140 caracteres, e permite também que o usuário siga páginas de conteúdos de outros usuários.

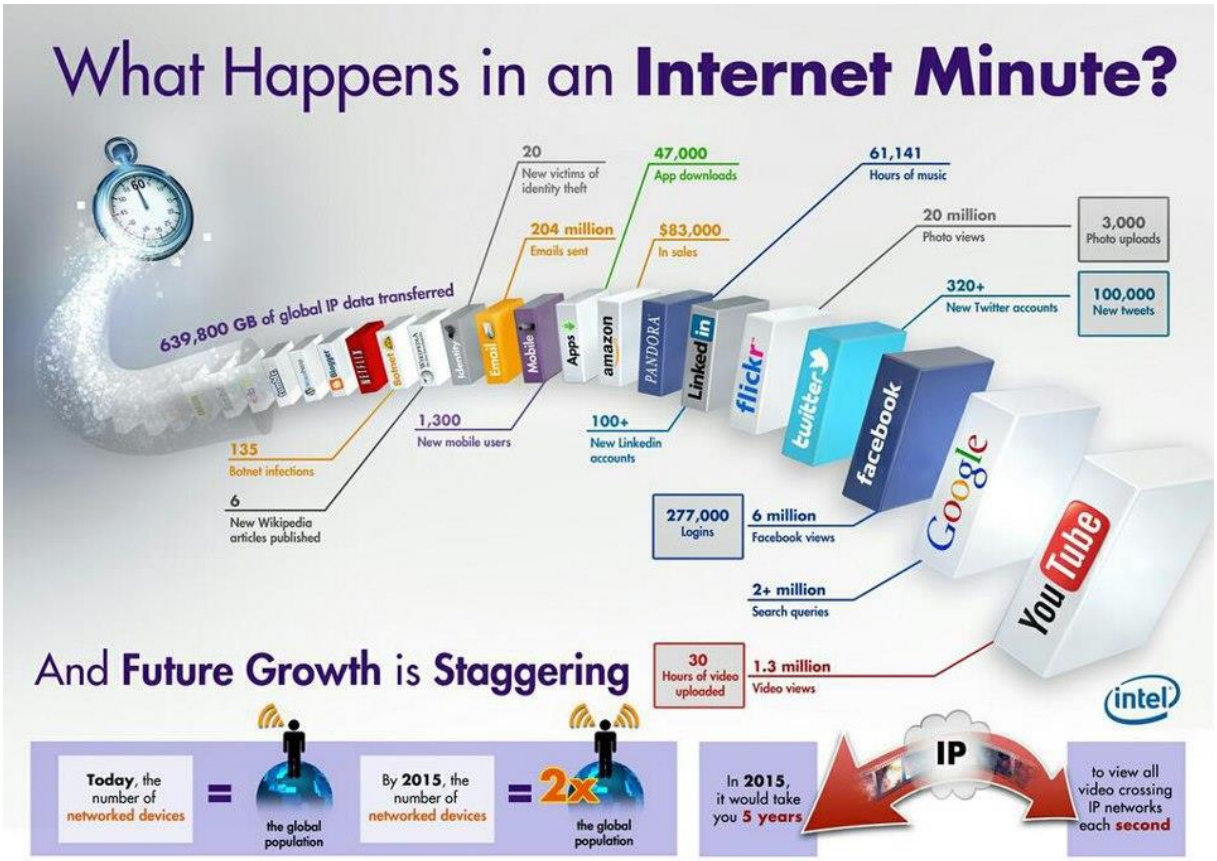
Com a crescente popularização das redes sociais e da Internet, surgiu a necessidade de explorar os dados gerados e extrair conhecimento. No caso do Twitter, a empresa responsável criou duas APIs (Application Programming Interface) que realizam a extração dos dados da rede social. Neste trabalho é utilizada a REST API[1]. A esses dados podem ser aplicadas técnicas de análise de sentimento para extrair conhecimento sobre o comportamento emocional dos usuários.

Um dos problemas que se encontra quando se realiza a extração de dados do Twitter é que ela é feita, normalmente, em todo o conteúdo de um usuário ou da rede social. As ferramentas existentes hoje não dispõem de uma opção para classificar os textos publicados em assuntos específicos.

Um exemplo simples da importância de uma aplicação que disponibiliza esses *tweets* em categorias específicas é a coleta dessas informações para gerar conhecimento. Esses *tweets* poderiam ser acessados de acordo com a categoria escolhida. Figueiredo e Garcia (2011) relatam que uma das atividades dos usuários do Twitter é a de buscar informações. Tendo isso em vista, este trabalho busca propor um modelo de classificação que utilize os dados publicados no Twitter para classifica-los automaticamente, como forma de disseminar informações importantes, separadas de acordo com categorias pré-definidas. As categorias utilizadas nesse trabalho são economia, esporte, religião, política e outros.

[1] https://dev.twitter.com/rest/public

Figura 1 – O que acontece na Internet em um minuto?



Fonte: IBM (2013)

O processo de criação de um modelo de classificação utilizando dados de redes sociais é um tópico já estudado por outros pesquisadores como o de (Gomide et al. 2011) para descobrir os focos de dengue no Brasil e o de (FILHO; LEITE; DA SILVA, 2014). O trabalho de (Carvalho et al. 2014) propos um classificador de *tweets* capaz de analisar o sentimento das mensagens em positivo, negativo, ambíguo e neutro nos dados referentes a Copa do Mundo 2014.

A próxima seção descreve os trabalhos relacionados que serviram de inspiração e base para o desenvolvimento deste trabalho, demonstrando as semelhanças e diferenças. Na Seção 3, serão informadas as fundamentações teóricas que contém os conceitos chave que serão utilizados neste trabalho. Na Seção 4, serão apresentados os procedimentos metodológicos, descrevendo minuciosamente todas as etapas do projeto. Na Seção 5, são apresentados os resultados obtidos. Na Seção 6, é apresentado uma discução sobre todo o trabalho. E por fim, na Seção 7, serão relatadas as considerações finais deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção será abordado os conceitos necessários para o entendimento deste trabalho. A seção 2.1 aborda os conceitos sobre Mineração de Textos, juntamente com as suas etapas para realizar esse processo. Na seção 2.2 será definido como funciona o Twitter, juntamente com a definição de algumas de suas funcionalidades. Na seção 2.3 abordará alguns conceitos de algoritmos de classificação, com ênfase no algoritmo de classificação Naive Bayes, que será abordado na subseção 2.3.1.

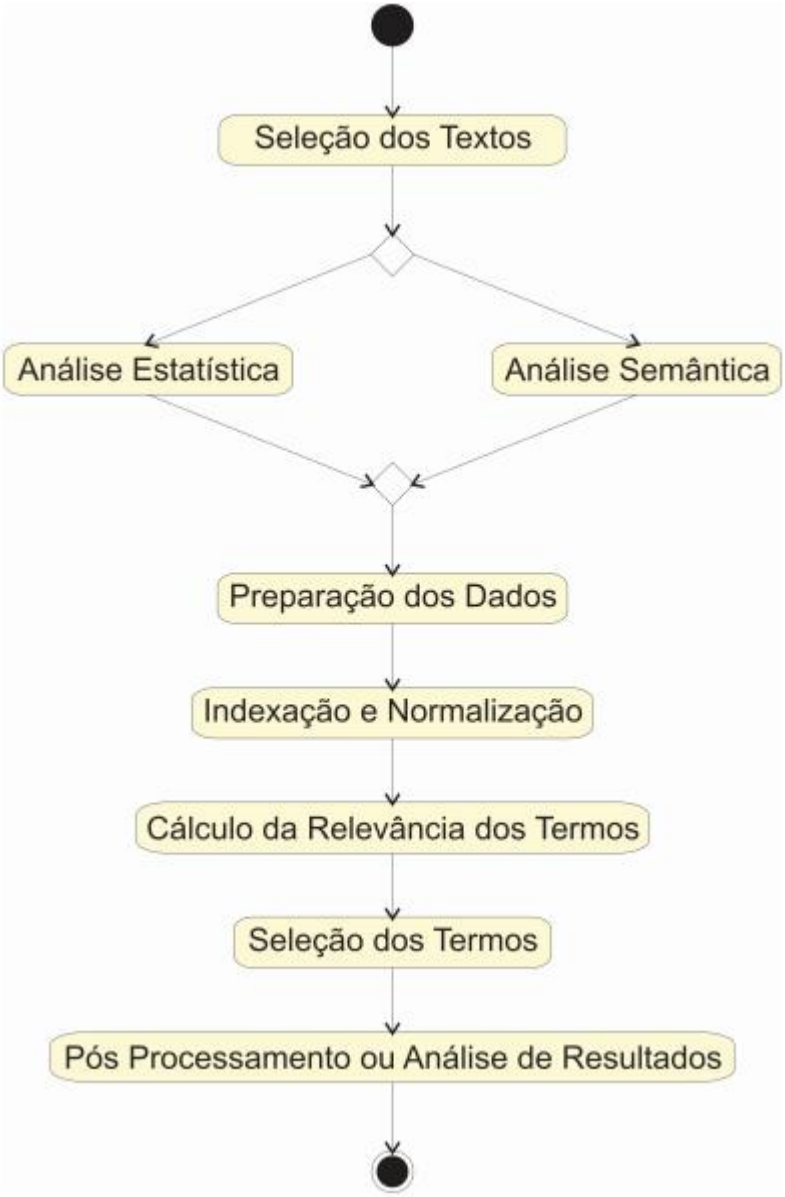
2.1 Mineração de Textos (*Text Mining*)

Mineração de textos, ou *text mining*, é definido por Morais e Ambrósio (2007) como uma técnica de análise e extração de conhecimento a partir de textos, frases ou apenas palavras, com o objetivo de identificar informações úteis e implícitas, contidas nos dados armazenados em formato não estruturado.

De acordo com Dorre, Gerstl e Seiffert (1999), o formato de armazenamento de dados não estruturados expressa uma vasta quantidade de informações. Por mais que a sua extração seja bastante codificada e complicada de se obter, elas não podem deixar de ser realizadas. Kao e Poteet (2010) citam que essas informações englobam tanto Recuperação de Informação quanto Classificação ou Clusterização de documentos textuais.

A prática da mineração de textos pode ser realizada em qualquer domínio que utilize textos, normalmente contidos em documentos, aplicando-se algoritmos computacionais para processar os textos e conseguir obter conhecimento contido no formato de dados não estruturados. Morais e Ambrósio (2007) relatam que a mineração de textos constitui de um processo que engloba várias etapas que são: seleção de documentos, definição do tipo de abordagem, preparação dos dados, indexação e normalização, cálculo da relevância dos termos, seleção dos termos e pós-processamento. Essas etapas são ilustradas na Figura 2.

Figura 2 – Etapas do Processo de Mineração de Textos.



Fonte: Morais e Ambrósio (2007).

2.2 Twitter

Barbosa et al (2012) definiram o Twitter como uma rede social que induz os usuários a compartilhar e expressar, de forma contínua, as suas opiniões e sentimentos de forma rápida e curta, com o objetivo de compartilhar conteúdos postados por usuários para a rede social em geral ou para os seus amigos o que o usuário em questão está sentindo sobre um determinado assunto.

O Twitter permite ao usuário criar conteúdos diversos, pois uma de suas funcionalidades é a de realizar uma pergunta simples e direta na página inicial do sistema: “O

que está acontecendo?". Com essa simples pergunta, o Twitter gera, diariamente, um grande volume de informações, mesmo os usuários enviando mensagens pela rede utilizando 140 caracteres no máximo. Essas mensagens são chamadas, na rede social, de *tweet*.

De acordo com Figueiredo e Garcia (2011), esse serviço possui características que facilitam a difusão de conteúdo na rede. Algumas delas são:

- **Retweets** - Permite que o usuário publique um determinado *tweet* existente em sua timeline, caso o mesmo tenha achado importante. Com isso, ele compartilha com os seus seguidores o *tweet*, sendo uma das principais ferramentas de difusão de conteúdo na rede social em questão.
- **Hashtags** - Esta funcionalidade permite o usuário citar em seu *tweet* utilizando o símbolo # seguido de um termo. Normalmente esse termo é uma citação de uma palavra, resumindo o conteúdo que será publicado. O Twitter gera um link para cada hashtag, levando o usuário, caso seja clicado, a buscar os últimos resultados dos usuários que utilizaram essa mesma hashtag.
- **Assuntos do Momento (*Trending Topics*)** – São uma lista em tempo real das frases mais publicadas no Twitter pelo mundo todo. Esse recurso tem como objetivo abranger todos os assuntos, mas é possível filtrar por países.

De acordo com Sousa (2012), o Twitter possui duas APIs: Twitter search API, que pode recuperar as postagens mais recentes, entre seis e nove dias antes da consulta, de usuários a partir de requisições HTTP[2], e ainda a Twitter Streaming API que permite buscas atualizadas em tempo real, mantendo uma conexão HTTP com o servidor do Twitter. O Twitter criou uma nova API que se chama REST API, onde cada requisição abre uma conexão HTTP e fecha assim que recebe o retorno dos dados. Para este trabalho foi utilizada a REST API para realizar a coleta das características que foram explicadas nesta seção, buscando tanto coletar *tweets* antigos como os *tweets* que estão sendo postados em tempo real.

2.3 Classificação

Os algoritmos de Classificação realizam a predição de categorias. Os mais conhecidos são árvore de decisão, redes bayesianas e os vizinhos mais próximos.

De acordo com Tan, Steinbach e Kumar (2005), os algoritmos de classificação são usados para delegar uma tarefa de atribuição de objetos a uma das categorias pré-definidas.

[2] http://pt.wikipedia.org/wiki/Hypertext_Transfer_Protocol

De acordo com Da Silva et al (2013), eles definem a classificação como uma técnica de mineração de dados que está na categoria de aprendizagem supervisionada.

Classificação de vizinhos mais próximos é um procedimento de decisão não paramétrico que classifica um novo objeto na categoria de seus vizinhos mais próximos (COVER e HART, 1967). No algoritmo de árvore decisão é gerado um modelo em formato de árvore, onde os vértices são atributos e as arestas são nomeadas com os valores desses atributos. Nas folhas estão as categorias pré-definidas. O algoritmo de classificação utilizado nesse trabalho é o algoritmo de Naive Bayes, que será explicado a seguir.

2.3.1 Naive Bayes

De acordo com Pang, Lee e Vaithyanathan (2002), três métodos bastante utilizados para a classificação de textos são Naive Bayes, Maximum Entropy e Support Vector Machines. Para este projeto foi utilizado o método de Naive Bayes, pois os trabalhos relacionados, que o utilizaram, mostraram resultados satisfatórios nas etapas que assemelham-se com este trabalho.

Thomas Bayes desenvolveu a técnica de Bayes em meados do seculo XVIII e é normalmente chamado de fórmula de probabilidade condicional de um determinado evento, muito utilizado em Machine Learning (JUNIOR, 2008).

De acordo com Singh e Husain (2013), o classificador Naive Bayes é baseado na probabilidade de instrução que lhe foi dado por Bayes. Este teorema fornece uma probabilidade condicional de ocorrência no evento E1 onde E2 já ocorreu. O caso contrário também pode ser calculado através da fórmula:

$$\boldsymbol{P}\,(\boldsymbol{E1}|\boldsymbol{E2}) \;=\; \frac{\boldsymbol{P}(\boldsymbol{E2}|\boldsymbol{E1})\;\boldsymbol{P}(\boldsymbol{E1})}{\boldsymbol{P}(\boldsymbol{E2})}$$

De acordo com Zhang (2004), na classificação, o objetivo da etapa de aprendizagem do algoritmo de Naive Bayes é construir um classificador utilizando um conjunto de dados de treino. O processo de cálculo das probabilidades é feito no conjunto de treino. Tipicamente, um exemplo aplicado ao contexto deste trabalho é: T é um *tweet* representado utilizando uma tupla de valores de atributos (x_1, x_2, ..., x_n), em que xi é o valor do atributo x_i. Cada valor xi é uma palavra ou token. Imagine que c é uma categoria, por exemplo esporte. Logo para o exemplo T = {x_1,x_2, .., x_n} ou T= {bola, jogo, campeonato, Ronaldo} teremos a formula:

$$\boldsymbol{P}\,(\boldsymbol{c}|\boldsymbol{T}) \;=\; \frac{\boldsymbol{P}(\boldsymbol{T}|\boldsymbol{c})\;\boldsymbol{P}(\boldsymbol{c})}{\boldsymbol{P}(\boldsymbol{T})}$$

O algoritmo Mahout Naive Bayes vêm em duas fases: aprendizado e aplicação. Durante a aprendizagem, um conjunto de vectores característicos é dado para o algoritmo, cada vector é marcado com a categoria que o pertence. Neste trabalho, verificamos que cada vector contém a categoria em que cada *tweet* pertence, logo teremos vectores diferentes com *tweets* em categorias diferentes.

3 TRABALHOS RELACIONADOS

A seguir, são apresentados os três principais trabalhos relacionados a este.

3.1 Tweetmining: Análise de opinião contida em textos extraídos do twitter

O trabalho de Sousa (2012) tem como principal objetivo desenvolver uma ferramenta que utilizando técnicas de mineração de textos é capaz de realizar um mapeamento emocional dos usuários do Twitter na cidade de Campinas-SP. Esta ferramenta tem a funcionalidade de polarizar as opiniões dos usuários, separando os *tweets* em neutros e opinativos. Os *tweets* opinativos são classificados em opiniões positivas e negativas. No trabalho de Sousa (2012), foi utilizado o algoritmo de classificação SVM (Support Vector Machines) para realizar a categorização de opiniões contidas em cada *tweet*. A extração de dados foi feita por meio da API de search do Twitter.

Analisar a opinião dos usuários não é um objetivo deste trabalho, pois o foco do projeto não é análise de sentimento e sim aplicar o algoritmo de classificação para filtrar os conteúdos em categorias específicas. Como semelhança é possível destacar a utilização de técnicas de classificação para minerar e separar os *tweets* em categorias pré-definidas.

3.2 Text Mining: Sentiment Analysis on News Classification.

O trabalho de Gomes, Neto e Henriques (2013) tem como objetivo construir um modelo que seja capaz de avaliar a polaridade dos títulos de notícias de economia, disponíveis nos endereços de RSS Feeds, tecnologia utilizada para realizar a recuperação de notícias *online* (WANNER et al. 2009). Os pesquisadores utilizaram o software SAS[1] para realizar a análise textual, e ainda *web creawlers*, que são programas que visitam sites e que automaticamente realizam a extração dos dados. Outro objetivo do trabalho foi apresentar um documento relatando como foi utilizado o processo de mineração de textos para as organizações portuguesas, facilitando os projetos futuros que utilizassem a mesma técnica.

O trabalho de Gomes, Neto e Henriques (2013) relata as vantagens e as desvantagens do uso da rede social entre empresa e cliente. Uma das principais desvantagens é o descontrole que as empresas têm sobre o que os clientes falam sobre a mesma. Porém, algumas empresas perceberam que poderiam se utilizar dessa situação para criar vantagens competitivas.

Criar um modelo de análise de sentimento, que avalie em categorias positivas, negativas e neutras não é um objetivo deste trabalho. Outro diferencial do trabalho de Gomes, Neto e Henriques (2013) é a utilização de um software, SAS, para realizar a análise textual. Neste projeto, será utilizada outra ferramenta *Mahout*[3]. A utilização de *web creawlers* para realizar a extração de dados também não será utilizada no desenvolvimento deste trabalho, pois para isto irá ser utilizado a API REST [4]do Twitter no processo de extração de dados. Como semelhança, é possível destacar a utilização de um algoritmo de classificação para categorizar os textos coletados.

3.2 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

O trabalho de Filho, Leite e Da Silva(2013) tem como objetivo principal analisar os sentimentos dos usuários do Twitter referentes aos acontecimentos da copa do mundo de 2015. Para isso, foi construido um modelo utilizando o algoritmo de Naive Bayes, capaz de avaliar a polaridade de cada *tweet*, classificando-os em positivo, negativo, neutro ou ambíguo. Os pesquisadores utilizaram a API REST do Twitter para realizar a coleta dos dados. Outro objetivo do trabalho, foi apresentar uma nuvem de palavras contendo as *hashtag's*[5] que foram mais utilizadas dentre os usuários em cada jogo da seleção brasileira.

Criar um modelo de análise de sentimento, que avalie em categorias positivas, negativas, neutras ou ambiguas não é um objetivo deste trabalho. Outro diferencial do trabalho de Filho, Leite e Da Silva (2014) é o formato da classificação dos *tweets*, neste trabalho iremos classifica-los em categorias pré-definidas. Como semelhança, é possível destacar a utilização de um algoritmo de classificação para categorizar os textos coletados. Também podemos destacar a utilização do algoritmo Naive Bayes para criar o modelo de classificação. A utilização de nuvens de palavras para destacar quais palavras estão sendo mais utilizadas nos *tweets* também é uma semelhança deste trabalho.

[3] http://mahout.apache.org/
[4] https://dev.twitter.com/rest/public
[5] http://pt.wikipedia.org/wiki/Hashtag

Tabela 1 – Comparação entre os trabalhos relacionados e este trabalho

Trabalho	**Fonte dos dados**	**Objetivo**	**Algoritmo de Classificação**	**Ferramenta escalável**
Sousa (2012)	Twitter	Análise de Sentimentos	SVM (Support Vector Machines)	Não utilizou
Gome, Neto e Henriques (2013)	*web crawlers*	Análise de Sentimentos	Naive Bayes	Não utilizou
Filho, Leite e Da Silva (2014)	Twitter	Análise de Sentimentos	Naive Bayes	Apache Mahout
Este trabalho	**Twitter**	**Categorização em notícias**	**Naive Bayes**	**Apache Mahout**

Fonte: criado pelo autor.

4 PROCEDIMENTOS

A fim de alcançar o objetivo deste trabalho de criar um modelo de classificação, com finalidade de categorizar os *tweets* coletados do Twitter e classifica-los automaticamente serão adotados alguns procedimentos para esta elaboração, são eles:

4.1 Coleta dos Dados

A coleta dos *tweets* foi o primeiro passo de execução deste trabalho. Para isto, foi criado um *script* em Python. O *script* tem como objetivo receber uma lista de valores, correspondentes às palavras ou hashtags que se relacionam a alguma categoria pré-definida. Com isso, serão retornados os *tweets* que possuem algumas dessas palavras. Cada *tweet* possui seus metadados: id, texto, coordenadas e data de criação. A API do Twitter permite delimitar a latitude, longitude e raio da pesquisa dos *tweets*, com isso, configuramos o *script* para realizar a coleta em um determinado raio que corresponde somente ao Brasil. Os *tweets* foram armazenados em um arquivo TSV (Tab-Separated Value) que separa cada campo do *tweet* por tabulação, foram armazenados 500 *tweets* por categoria.

4.2 Limpeza dos Dados

Para realizar a limpeza dos *tweets* foi utilizado outro script em Python para realizar o pré-processamento dos dados. Esse *script* recebe dois parâmetros, o nome do arquivo TSV de entrada e o nome do arquivo TSV que será gravado quando a limpeza for feita.

O objetivo principal desta limpeza é remover e organizar conteúdos irrelevantes para o processo de classificação dos tweets. Esse processo envole as seguintes etapas:

- **Remoção de acentos**: Como a coleta foi realizada através da rede social, a informalidade entre os usuários é muito grande. Com isso, existem usuários que preferem digitar palavras com ou sem acentos. Essas palavras possuem o mesmo significado, mas para o algoritmo de classificação esses *tweets* apresentam palavras diferentes. Por causa deste motivo, esta etapa foi realizada para evitar esse problema.
- **Remoção de pontuação**: Pelo mesmo motivo anterior, as palavras que são seguidas ou precedidas de alguma pontuação são interpretadas como sendo palavras diferentes. Por isso, é necessário retirar a pontuação dos textos coletados.

As próximas etapas são referentes ao processamento de linguagem natural utilizado neste trabalho. Para isso, foram realizadas com a utilização da plataforma NLTK[6] (Natural Language ToolKit), escrita em Python, que trabalha com dados de linguagem humana.

- ***Remoção de Stop Words***[7]: Remoção de palavras que não possuem relevância para os resultados da classificação de textos, como artigos, preposições, pronomes, entre outros.
- **Tokenização:** Foi necessário separar, cada palavra de cada *tweet,* e construir um vetor de *tokens* a partir das mesmas.
- **Remoção de Sufixos (Redução ao radical)**: Foi necessário remover os sufixos das palavras para descobrir que elas podem se equivaler em significado, mesmo escritas com desinências verbais e nominais diferentes. Como as palavras quero e queremos, por exemplo.

Os *tweets* repetidos e os que não foram escritos em português também foram descartados. Os *retweets* foram removidos, pois geram uma grande quantidade de textos replicados na base coletada. Para isso, foi utilizado outro *script* que utiliza, também, a biblioteca NLTK para identificar a linguagem do corpo do *tweet*, com isso, ele remove os *tweets* que são identificados em outra linguagem que não seja o português.

4.3 Definição das Categorias

Nesta etapa do projeto foram definidas quais categorias considerar para alocar os *tweets*. As categorias escolhidas neste trabalho são as categorias que estão mais presentes em sites de notícias. Os *tweets* coletados passaram por um processo de categorização manual, realizada pelo autor deste trabalho, para que fosse construído e testado o modelo de classificação com o objetivo de automatizar este processo. As categorias utilizadas foram economia, esportes, religião, política e outros.

O processo de categorização foi realizado manualmente, por não existir uma ferramenta que os categorize automaticamente. No entanto, por meio do modelo gerado neste trabalho é possível categorizar automaticamente novos *tweets* em assuntos como economia, esporte, religião, política e outros.

[6] nltk.org
[7] http://pt.wikipedia.org/wiki/Palavra_vazia

4.4 Mineração de Textos

Para realizar esta etapa, foi utilizado o algoritmo de classificação de textos Naive Bayes do Apache Mahout para gerar o modelo de classificação de *tweets*. Esta ferramenta foi escolhida por possuir código aberto, ser uma ferramenta escalável e mostrar resultados satisfatórios nos trabalhos relacionados que mais se assemelham a este trabalho.

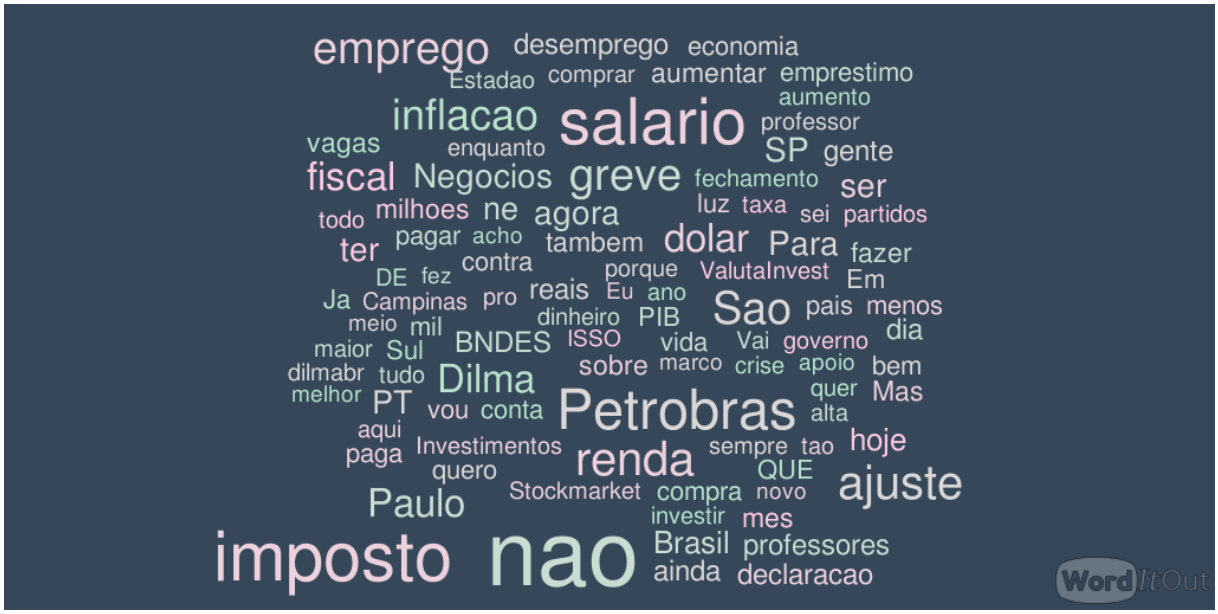
Para isso, foi utilizado um conjunto de *tweets*, coletados e pré-processados para treino, isto é, os que auxiliam no processo de criação do modelo de classificação. Como informado anteriormente, estes *tweets* foram classificados manualmente em alguma das cinco categorias pré-definidas. Um novo conjunto de *tweets*, também coletados (125 novos *tweets*) e pré-processados conforme visto na Seção 4.2, foi utilizado para testar o modelo. Assim, é possível avaliar o modelo gerado a partir do conjunto de treino, a fim de verificar a acurácia do modelo de classificação, para validar o modelo final que foi gerado.

A fim de utilizar o modelo e classificar novos *tweets* automaticamente, foi realizada uma nova coleta de *tweets* utilizando o mesmo processo da Seção 4.1, além de verificar se a acurácia do modelo satisfaz a realidade. Para isso, foi verificado se a acurácia de classificação em cada categoria era semelhante a acurácia de classificação do conjunto de teste. Neste trabalho, o modelo foi capaz de categorizar automaticamente os novos *tweets*, no entanto apenas para algumas categorias ele apresentou acurácia conforme indicado na fase de teste. No conjunto de teste é possível verificar a porcentagem de acerto em cada categoria do mogelo de classificação, com isso foi relizado uma comparação do conjunto de teste com a nova coleta, afim de verificar se a acurácia era semelhante.

4.5 Criação das Nuvens de Palavras

Por fim, este trabalho apresenta as nuvens de palavras de cada categoria, que são um recurso gráfico para descrever a frequência de cada palavra nos *tweets*. As palavras que são mais utilizadas dentre os usuários, terão uma relevância maior na imagem gerada. Este passo tem como objetivo informar quais palavras estão sendo mais utilizadas, entre os usuários, em diferentes categorias. Esse recurso permite verificar se alguma palavra está sendo utilizada para classificar categorias diferentes. Neste trabalho, foi possível notar a presença de palavras iguais em diferentes categorias, com isso pode ser explicado o motivo do algoritmo de classificação classificar incorretamente algum *tweet*. Um exemplo de nuvem de palavras de economia é apresentado na Figura 3.

Figura 3 – Exemplo de núvem de palavras de economia



Fonte: elaborada pelo autor.

5 DESENVOLVIMENTO/RESULTADOS

Esta seção tem como objetivo relatar, detalhadamente, a cerca da realização dos procedimentos citados anteriormente. A seção é composta de três subseções. A Seção 5.1 retrata a primeira etapa deste projeto, onde foi realizada as coletas dos dados do Twitter. A Seção 5.2 descreve detalhatadamente como foi construído o modelo de classificação e como foi testado. A Seção 5.3 tem como objetivo principal comparar as palavras mais utilizadas nos *tweets* de categorias diferentes, afim de explicar o motivo do modelo classificatório classificar alguns *tweets* incorretamente.

5.1 Coleta de Dados

No processo de coleta de dados, foram coletados 41.153 *tweets* de dezembro de 2014 até maio de 2015. Para isso, foi utilizado um conjunto de palavras relacionadas a cada categoria pré-definida. Este conjunto de palavras foi retirado dos principais sites de notícias do Brasil, selecionando algumas palavras através das notícias de cada categoria específica. Para cada coleta, seus valores foram salvos em um arquivo TSV separado por tabulação, essa estratégia foram utilizadas para que os valores fiquem organizados em colunas diferentes. Os *tweets* foram separados em arquivos diferentes, como foram utilizadas cinco categorias, cinco arquivos TSV foram criados. Como a coleta foi realizada somente com *tweets* brasileiros, só foi utilizado um metadado para os *tweets*, que foi a mensagem do *tweet*.

Tabela 2 – Quantidade de *tweets* coletados

Descrição	**Quantidade**
Tweets	41.153
Tweets pré-processados	16.780
Tweets para criação do modelo	2.500
Tweets conjunto de treino	1.722 (70%)
Tweets conjunto de teste	778 (30%)
Nova coleta	125

Fonte: criado pelo autor.

5.1.1 Pré-processamento

Após a fase de coleta de dados, foi criado um script[9] para que os *tweets* fossem pré-processados conforme explicado na Seção de 4.2 deste trabalho.

Tabela 3 – Étapas do Pré-processamento

Étapas	**Exemplo**
Tweet original	Tudo piora. Mais imposto, inflação, corrupção menos escola, menos hospitais.
Remoção de acentos	Tudo piora. Mais imposto, inflacao, corrupcao menos escola, menos hospitais.
Remoção de pontuação	Tudo piora Mais imposto inflacao corrupcao menos escola menos hospitais
Texto em minúsculo	tudo piora mais imposto inflacao corrupcao menos escola menos hospitais
Remoção de Stopwords	tudo piora imposto inflacao corrupcao menos escola menos hospitais
Remoção de sufixo (*tweet* final)	tud pior impost inflaca corrupca menos escol menos hospit

Fonte: criado pelo autor.

5.2 Desenvolvimento e Verificação da Acurácia do Modelo de Classificação

Após a realização da filtragem dos textos coletados, foram selecionados 2.500 *tweets* para gerar o modelo de classificação. Para isso, foi necessário juntar os arquivos que foram coletados. Foi criado um TSV único com todas as cinco coletas, com esse arquivo foi criado o modelo de classificação. Os *tweets* deste arquivo foram classificados manualmente em cinco categorias (economia, esporte, outros, religião e política). Em sequida, esse conjunto foi dividido em conjunto de treino totalizando 69% dos *tweets* coletados. Foram selecionados 1.722 *tweets* randomicamente para criar o conjunto de treino e 31% dos *tweets* (778) selecionados randomicamente para criar o conjunto de teste. As 5 categorias estão uniformemente distruibuídas nos 2.500 *tweets*

Após utilizar o conjunto de treino para gerar o modelo, foi verificado a acurácia de classificação do mesmo. O modelo instanciou 1.681 *tweets* corretamente totalizando 97.619% dos *tweets*, e classificou 41 *tweets* incorretamente totalizando 2.381%, utilizando o conjunto de treino contendo 1.722 *tweets* para cria-lo.

[9] https://github.com/LucasES/ProcessamentoLinguagemNatural

Tabela 4 – Conjunto de treino

Descrição	**Quantidade**
Tweets Classificados Corretamente	1681 (97.619%)
Tweets Classificados Incorretamente	41 (2.381%)
Total de *tweets*	1722
Acurácia	97.619%

Fonte: criado pelo autor.

Os *tweets* utilizados para criar o conjunto de treino foram divididos aleatoriamente entre as categorias pré-definidas. Com isso podemos verificar na Tabela 5 que alguns *tweets* foram classificados nas categorias correspondentes ou incorretamente.

Tabela 5 – Quantidade de *tweets* classificados em cada categoria

Categoria	**Esporte**	**Economia**	**Religião**	**Política**	**Outros**
Esporte	332	0	1	0	0
Economia	1	327	1	8	3
Religião	0	0	354	0	0
Política	0	11	1	332	1
Outros	4	7	0	3	336

Fonte: criado pelo autor.

Após criar o modelo utilizando o conjunto de treino, foi verificada a acurácia de classificação do mesmo utilizando os *tweets* que foram separados em conjunto de teste. O modelo apresentou uma taxa de 83.6761% de precisão, utilizando o conjunto de treino contendo 778 *tweets* para cria-lo. A imagem 5 ilustra a precisão do conjunto de teste:

Tabela 6 – Conjunto de teste

Descrição	**Quantidade**
Tweets Classificados Corretamente	651 (83.676%)
Tweets Classificados Incorretamente	127 (16.323%)
Total de *tweets*	778
Acurácia	83.6761%

Fonte: criado pelo autor.

Os *tweets* utilizados para criar o conjunto de teste foram divididos aleatoriamente entre as categorias pré-definidas. Com isso podemos verificar na Tabela 7 que alguns *tweets* foram classificados nas categorias correspondentes ou incorretamente.

Tabela 7 – Quantidade de *tweets* classificados em cada categoria

Categoria	**Esporte**	**Economia**	**Religião**	**Política**	**Outros**
Esporte	160	2	0	1	4
Economia	5	117	3	18	17
Religião	0	0	142	2	2
Política	1	18	7	121	8
Outros	1	18	12	8	111

Fonte: criado pelo autor.

Após o teste com 778 *tweets*, foi realizada uma nova coleta com 125 novos *tweets*. Para isso, foram coletados 25 *tweets* de cada categoria, passando pelo processo de categorização manual e pré-processamento para verificar a taxa de acurácia dessas novas instâncias de dados.

Dos 25 *tweets* referentes a categoria de esportes obtivemos 100% de acerto. Ou seja, todas as 25 instâncias foram classificadas na categoria correta.

Foram selecionados 25 *tweets* de política para serem testados. O modelo gerado classificou 7 instâncias incorretamente, dessas instâncias tivemos um *tweet* classificado na categoria outros e 6 *tweets* classificado na categoria economia. Ou seja, para esta categoria a taxa de acurácia do modelo de classificação foi de 72%. Podemos concluir que, a taxa de instâncias classifificadas em economia, que eram para serem classificadas em política, foi bastante alta, mas podemos verificar o motivo disto através de um estudo realizado através da comparação das nuvens de palavras de economia e de política. Através disso, podemos verificar que existem palavras, entre essas duas categorias, que são utilizadas pelos usuários para referenciarem a diferentes assuntos, tais palavras são: professores, petrobras, governo, dilma entre outras que são muito utilizadas no contexto política e no contexto economia.

Foram selecionados, novamente, 25 *tweets* de religião para serem testados. O modelo gerado classificou 6 instâncias incorretamente, dessas instâncias tivemos um *tweet* classificado na categoria outros, 2 *tweets* classificados na categoria de esportes e 3 *tweets*

classificados na categoria política. Ou seja, para esta categoria a taxa de acurácia do modelo de classificação foi de 76%.

Foram selecionados, novamente, 25 *tweets* de economia para serem testados. O modelo gerado classificou 11 instâncias incorretamente, dessas instâncias tivemos 7 *tweets* classificados na categoria política, 3 *tweets* classificados na categoria outros e 1 *tweet* classificado na categoria religião. Ou seja, para esta categoria a taxa de acurácia do modelo de classificação foi de 56%. Pelo mesmo motivo que foi explicado anteriormente, existem palavras entre as categorias economia e política que são bastante usadas entre os usuários. Por este motivo, o modelo de classificação acaba classificando essas instancias incorretamente.

Foram selecionados, novamente, 25 *tweets* da categoria outros para serem testados. O modelo gerado classificou 20 instâncias incorretamente, dessas instâncias tivemos 10 *tweets* classificados na categoria economia, 8 *tweets* classificados na categoria política e 2 *tweets* classificados na categoria esporte. Ou seja, para esta categoria a taxa de acurácia do modelo de classificação foi de 20%. A baixa taxa de acerto desta categoria é pelo fato de que é uma categoria que envolve um contexto geral de notícias. Ou seja, como esta categoria aborda assuntos de todas as outras o modelo de classificação acaba categorizando as instancias incorretamente. Para resolver este problema deve-se criar um modelo de treino com mais instâncias na categoria outros, para que o novo modelo gerado esteja com uma taxa de aprendizagem mais alta do que o modelo que foi criado neste trabalho.

5.3 Criação da Nuvem de Palavras utilizando o website WordItOut

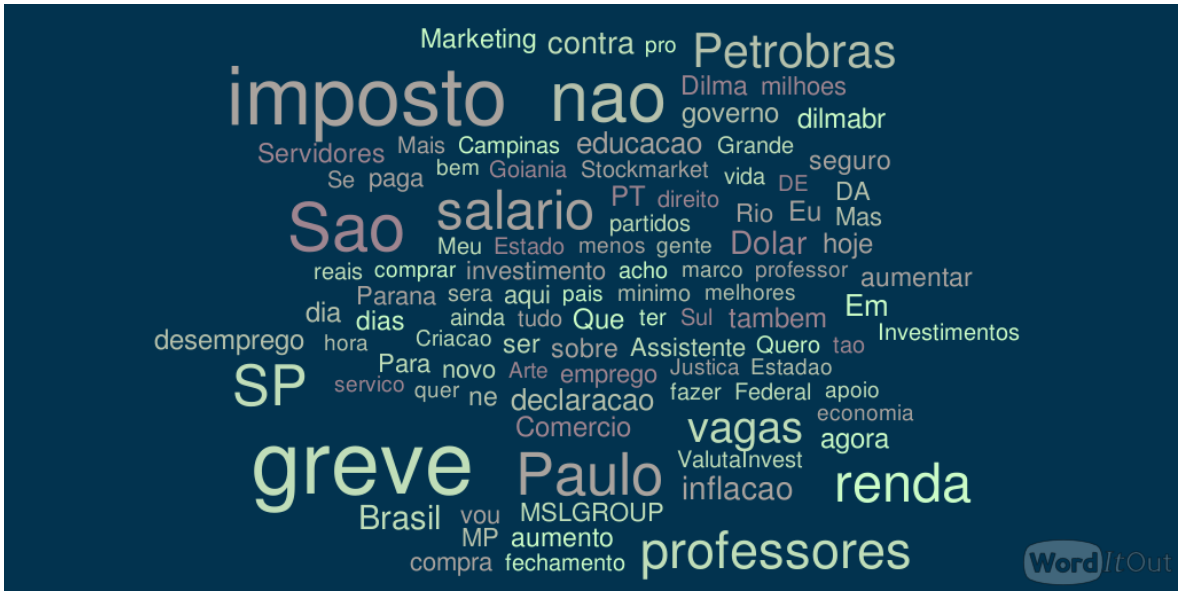
Foram criadas as nuvens de palavras de cada categoria pré-definida. As nuvens de palavras tem como objetivo ilustrar quais palavras foram mais utilizadas em cada categoria. Um dos objetivos da geração das nuvens de palavras neste trabalho é verificar se existem palavras que foram bastante utilizadas em diferentes categorias, caso existam, será um dos motivos da baixa da acurácia do modelo de classificação.

5.3.1 Conjunto de treino

5.3.1.2 Economia

A nuvem de palavras desta categoria foi criada através de um arquivo TSV contendo os *tweets* de economia que foram utilizados para criar o modelo de treino. Este arquivo contém 340 *tweets* que foram classificados manualmente na categoria economia.

Figura 4 – Nuvem de palavras de economia



Fonte: elaborada pelo autor.

A categoria economia contem algumas palavras que estão em alta no momento (Abril 2015). Dentre elas podemos citar: greve, professores, imposto, renda entre outras. Essas palavras são as que os usuários do Twitter estão utilizando com mais frequência, de acordo com os acontecimentos que estamos tendo neste momento. Por exemplo as palavras, greve e professores, são referentes as greves que estão ocorrendo no sul do Brasil.

5.3.1.2 Esportes

Para esta categoria foram utilizados 333 *tweets*, coletados e pré-processados, utilizados para gerar o modelo de treino. Esses *tweets* foram classificados manualmente, pelo autor, na categoria esportes.

Figura 5 – Nuvem de palavras de esportes

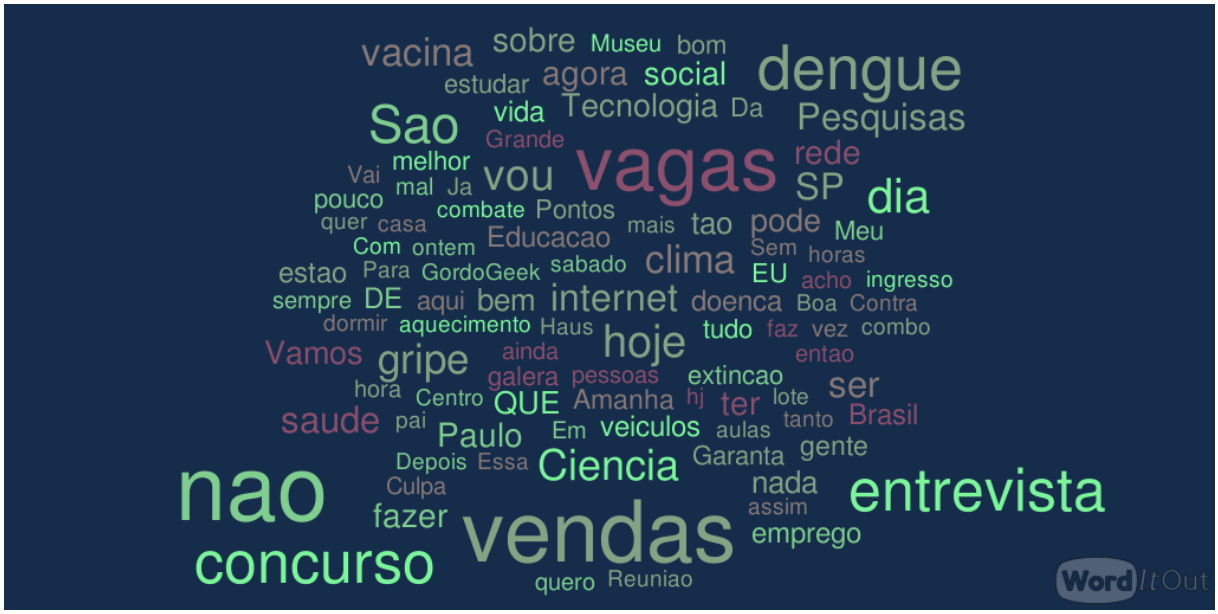
Fonte: elaborada pelo autor.

É possível verificar que algumas palavras têm uma maior taxa de utilização entre os usuários do Twitter. Dentre elas podemos citar: esporte, campeonato, MMA, Brasileiro. É possível concluir que, a maioria dos usuários do Twitter que publicam sobre esporte utilizam essa mesma palavra em seu tweet . Notouse que a palavra surf apareceu na nuvem de palavras. Podemos conciliar isso com o brasileiro que venceu o campeonato mundial de surf de 2014 (Gabriel Medina).

5.3.3 Outros

Para esta categoria foram utilizados 350 *tweets*, coletados e pré-processados, utilizados para gerar o modelo de treino. Esses *tweets* foram classificados manualmente, pelo autor, na categoria outros. Os *tweets* coletados para está categoria foram coletados através de um dicionário de palavras contendo palavras de diversas categorias de notícias.

Figura 6 – Nuvem de palavras da categoria outros



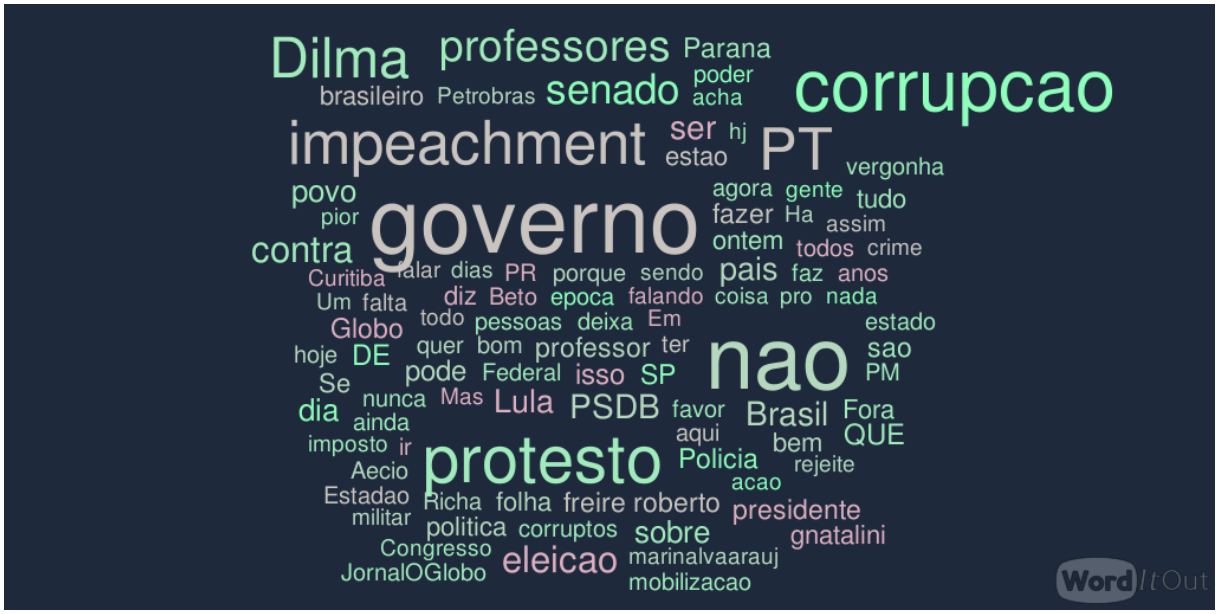
Fonte: elaborada pelo autor.

A ilustração mostra que algumas palavras estão em alta no momento da coleta da categoria outros. Algumas delas são: vagas, vendas, concurso, entrevista, dengue. Como essas palavras são as mais utilizadas nesta categoria, o modelo de classificação terá uma maior facilidade para classificar os novos *tweets* que contenham essas palavras.

5.3.4 Política

Para esta categoria foram utilizados 345 *tweets*, coletados e pré-processados, utilizados para gerar o modelo de treino. Esses *tweets* foram classificados manualmente, pelo autor, na categoria política.

Figura 7 – Nuvem de palavras de política



Fonte: elaborada pelo autor.

Podemos verificar que algumas palavras têm uma maior taxa de utilização entre os usuários do Twitter. Dentre elas podemos citar: governo, corrupcao, protesto, dilma, professores. A partir desta nuvem de palavras, podemos concluir que o modelo poderá classificar incorretamente alguns *tweets*, pois como verificado anteriormente na nuvem de palavras de economia temos palavras que aparecem, frequentemente, nessas duas categorias. Por exemplo: professores, PT, SP. Logo os tweets que conter algumas dessas palavras foram classificados em alguma dessas duas categorias.

5.3.5 Religião

Para esta categoria foram utilizados 354 *tweets*, coletados e pré-processados, utilizados para gerar o modelo de treino. Esses tweets foram classificados manualmente, pelo autor, na categoria política.

Figura 8 – Nuvem de palavras de religião

Fonte: elaborada pelo autor.

Podemos verificar que a categoria religião possui algumas palavras que são bastante utilizadas entre os usuários do Twitter. Dentre elas podemos citar: Cristo, Deus, Gospel, Igreja, Religiao, Espirito. Logo podemos verificar que, essas palavras que são frequentemente utilizadas entre os usuários, terão uma probabilidade maior de novos *tweets* serem classificados corretamente na categoria religião.

6 DISCUSSÃO

Os resultados apresentados ao final da criação do modelo de classificação comprovaram que é possível classificar novos *tweets* em categorias pré-definidas escolhidas pelo pesquisador. A coleta foi realizada em todos os estados brasileiros e com diferentes usuários.

Podemos verificar que existem diferentes tipos de usuários na rede social do Twitter que buscam e postam assuntos diversos. A possibilidade de filtrar as postagens em categorias pode ajudar os usuários que estão a procura de algum assunto específico.

Assim, a proposta de criação do modelo de classificação contribuirá com os usuários ou pesquisadores que necessitem utilizar desses dados para realizar um novo estudo ou divulgar noticias através da utilização da rede social.

Isso mostra que, embora exista um esforço para a criação de um modelo de classificação, onde a categorização esteja com uma alta probabilidade acertos, os resultados obtidos podem satisfazer a diferentes públicos.

Podemos verificar que os assuntos do momento mudam a cada momento. Com isso, o modelo de classificação deverá ser realimentado com novos dados caso não possua nenhum dado sobre alguma nova notícia que poderá acontecer no futuro.

A classificação automática dos *tweets* gera uma núvem de palavras semelhante a Seção 5.3, pois o algoritmo organiza os *tweets* que passaram pelo modelo de classificação em cada categoria pré-definida. Com isso, para criar a nuvem de palavras desses dados deve-se realizar o mesmo processo da Seção 5.3.

7 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o processo de Mineração de Textos realizado a partir dos textos coletados, pré-processados do Twitter (*tweets*). Também mostrou como foi criado o modelo de classificação desses textos através da categorização dos *tweets* em categorias pré-definidas. Os *tweets* coletados neste trabalho tiveram que ser pré-processados para remover todas as palavras que não possuem relevância para gerar o modelo. Inicialmente, foram descartados os *tweets* repetidos e os que não eram em português. A etapa que necessitou de mais tempo foi a categorização dos *tweets*, pois este processo foi feito manualmente lendo um *tweet* de cada vez, verificando se o mesmo se encaixa na categoria escolhida. Como foram escolhidas cinco categorias, este processo necessitou de uma quantidade de tempo maior. Este processo é altamente importante, pois para criar o modelo de classificação, a partir do conjunto de treino, deve-se ter um conjunto de dados categorizados corretamente.

Com isso, o modelo gerado neste trabalho, permitiu mostrar que é possível categorizar *tweets* em categorias pré-definidas com uma precisão bastante satisfatória utilizando o algoritmo de classificação Naive Bayes, verificando as novas instâncias (*tweets*) com o que foi gerado através do conjunto de treino. Como a categoria outros abrange um contexto geral, o modelo gerado não demonstrou uma classificação totalmente precisa, para aprimorar isto deve-se criar um novo modelo de classificação com novos *tweets* classificados corretamente nesta categoria.

Como trabalho futuro deve-se adeguar o modelo de classificação a notícias de diferentes. Para isso, deve-se realimentar o modelo final para que o mesmo possa continuar classificando as novas notícias.

Outro trabalho futuro será disponibilizar em um website esses *tweets* categorizados automaticamente para o público geral. Com isso o objetivo deste trabalho futuro é criar um portal de notícias utilizando os dados do Twitter como fonte.

O trabalho de Filho, Leite e Da Silva(2014), serviu de base para a construção deste trabalho. Os processos de coleta, pré-processamento e criação do modelo de classificação foram bastantes semelhantes. A utilização do Mahout com Naive Bayes mostrou-se uma ótima ferramente para a categorização de textos.

Este trabalho contribuiu com os usuários que desejam buscar informações relevantes através das redes sociais. Os usuários que utilizarem este modelo, poderão realizar buscas através dos dados do Twitter filtrando-os em cinco categorias diferentes: economia, esportes, política, religião e outros. Por exemplo, imagine usuários que estejam buscando uma

vaga em algum emprego através do Twitter, com o modelo de classificação gerado neste trabalho, esses *tweets* poderiam estar listados e disponibilizados nas categorias outros.

Para ilustrar o motivo, do modelo classificar novas instâncias incorretamente, foram utilizadas nuvens de palavras. Com isso, foram obtidas as palavras que foram mais utitilizadas entre os usuários do Twitter. A partir disto, foi comprovado que os usuários utilizam algumas palavras semelhantes em contextos (categorias) diferentes. Logo se a frequência de utilização dessas palavras for alta, em categorias diferentes, o modelo classificará os novos *tweets* em algumas dessas categorias.

A elaboração deste trabalho serviu como base para contribuir no 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[10].

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

REFERÊNCIAS

BARBOSA, G. A. R.; JR, W. M.; PRATES, R. O.; SILVA, I. S.; VELOSO, A; ZAKI, M. J. Characterizing the Effectiveness of Twitter Hashtags to Detect and Track Online Population Sentiment. Local: Texas – USA, 2012. Disponível em: < http://www.cs.rpi.edu/~zaki/PaperDir/CHI12.pdf >. Acesso em: 21 mar. 2014.

COVER, T.; HART, P. Nearest neighbor Pattern Classification. Information Theory, IEEE Transactions. Local: Piscataway - USA, 1967

DA SILVA, T. L.; SOUSA, F. R.; DE MACÊDO, J. A. F.; MACHADO, J. C.; CAVALCANTE, A. A. Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem. **28° Simpósio Brasileiro de Banco de Dados**. Local: Recife – PE, 2013.

DORRE, J.; Gerstl GERSTL, P.; SEIFFERT, R. Text mining: finding nuggets in mountains of textual data. **Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 398–401)**. Local: New York, NY. 1999.

FIGUEIREDO, K. S.; GARCIA, A. C. B. Uma Investigação Sobre os "Assuntos do Momento" e a Discussão de Notícias no Serviço de Microblogging Twitter. **VIII Simpósio Brasileiro de Sistemas Colaborativos**, Local: Paraty – RJ, 2011.

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.

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

GOMES, H.; NETO, M. C.; HENRIQUES, R**.** Text Mining: Sentiment Analysis on News Classification. Universidade Nova de Lisboa. Local: Lisboa – PT, 2013. Dísponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6615822>. Acesso em: 22 mar. 2014.

GOMIDE, J.; VELOSO, A.; MEIRA Jr, W.; ALMEIDA, V.; BENEVENUTO, F., FERRAZ, F.; TEIXEIRA, M. Dengue Surveillance Based on a Computational Model of Spatio-Temporal Locality of Twitter**. Proceedings of the 3rd International Web Science Conference**. Local: New York – USA, 2011.

JUNIOR, J. R. C. Desenvolvimento de uma Metodologia para Mineração de Textos. **Pontífica Universidade Católica do Rio de Janeiro**. Rio de Janeiro - RJ. 2008.

KAO, A.; POTEET, S. R. **Natural Language Processing and Text Mining** (1st ed.). Local: Springer, 2010.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de Textos. **Technical Report, Universidade Federal de Goiás**. Local: Goiás – GO. 2007. Disponível em:

<http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em: 16 mar. 2014.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. **Association for Computational Linguistics**, PA - USA 2002.

QUINLAN, J. R. Miniboosting Decision Trees. **Journal of Artificial Intelligence Research**. Local: Sydney - AUS, 1998

SINGH, P. K.; HUSAIN, M. S. Analytical Study of Feature Extraction Techniques in Opinion Mining. Local: **Computer Science**, 2013.

SOUSA, G. L. S. Tweetmining: Análise de Opinião Contida em Textos Extraídos do Twitter. Universidade Federal de Lavras, Local: Lavras – MG, 2012. Disponível em: Disponível em: <http://www.bsi.ufla.br/wp-content/uploads/2013/09/TWEETMINING-AN%C3%81LISE-DE-OPINI%C3%83O-CONTIDA-EM-TEXTOS-EXTRA%C3%8DDOS-DO-TWITTER-.pdf>. Acesso em: 02 mar. 2014.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 1° Edition. Local: Addison-Wesley, 2005.

WANNER, F.; ROHRDANTZ, C.; MANSMANN, F.; OELKE, D.; KEIM, D. A. Visual Sentiment Analysis of RSS News Feeds Featuring the us Presidential Election in 2008. **Workshop on Visual Interfaces to the Social and the Semantic Web**. Local: University of Konstanz - Alemanha, 2009.

WU, X.; ZHU, X; WU, G.-Q.; DING, W. Data Mining with Big Data. **IEEE Transactions on Knowledge and Data Engineering**. Local: Hefei - China, 2014.

ZHANG, Harry. The optimality of naive Bayes. **AA**, v. 1, n. 2, p. 3, 2004.