UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ZARATHON LOPES VIANA

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTO UTILIZANDO TWEETS REFERENTES ÀS ELEIÇÕES PRESIDENCIAIS 2014

QUIXADÁ
2014

ZARATHON LOPES VIANA

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTO UTILIZANDO TWEETS REFERENTES ÀS ELEIÇÕES PRESIDENCIAIS 2014

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

Área de concentração: Computação, Mineração de Dados

Orientadora Profª. MSc. Ticiana Linhares Coelho da Silva

QUIXADÁ
2014

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

V667m Viana, Zarathon Lopes
Mineração de textos: análise de sentimentos utilizando Tweets referentes às eleições presidenciais 2014 / Zarathon Lopes Viana. – 2014.
32 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, 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

ZARATHON LOPES VIANA

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTO UTILIZANDO TWEETS REFERENTES ÀS ELEIÇÕES PRESIDENCIAIS 2014

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

Área de concentração: Computação, Mineração de Dados

Aprovado em: _____ / dezembro / 2014.

BANCA EXAMINADORA

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

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

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

Aos meus pais, amigos e em especial a minha orientadora que aceitou o desafio de orientar um desorientado...

"Que os vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível."
(Charles Chaplin)

RESUMO

O trabalho a seguir tenta auferir o sentimento dos usuários da rede social Twitter em relação a eleição dos candidatos à presidência do Brasil no ano de 2014 no que se refere o primeiro turno desta. Utilizam-se técnicas de mineração de dados e classificação dos mesmos para extrair o que os usuários sentiam com relação a cada candidato através de suas mensagens. Positivo, negativo, neutro e ambíguo foram os parâmetros utilizados para classificar os sentimentos das mensagens.

Palavras chave: Mineração de Dados. Análise de Sentimentos. Twitter.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Vivemos em um mundo tecnologicamente social. É um fato que hoje por todo lado se escuta falar das redes sociais e como as mesmas estão mudando os hábitos do nosso dia-a-dia. São milhares de informações trocadas por minuto, com uma velocidade espantosa, seja uma notícia, seja um acontecimento nos mais remotos lugares da terra, qualquer informação, em questão de minutos, está difundida para o mundo.

Estamos conectados, direta ou indiretamente, nesse fenômeno mundial que molda nossos conceitos, que nos faz refletir sobre os acontecimentos. O Twitter, uma das 5 maiores redes sociais utilizadas no Brasil[1], é baseada em texto (microblog), onde o usuário interage através de mensagens com 140 caracteres. No Brasil, o número de usuários do Twitter ultrapassa a casa dos 210 milhões de usuários ativos. Para se ter uma ideia da quantidade de informação gerada por essa rede social, no carnaval de 2014 foram mais de 4,9 milhões de tuites (mensagem que o usuário publica) sobre a folia nas ruas e cidades do país[2].

Analisar volume de informações sendo gerada a cada instante pode nos levar a extrair um conhecimento coletivo, ou melhor, poderíamos utilizar isso no nosso cotidiano. A Mineração de Dados, nesse caso, Mineração de Texto, nos apresenta um conjunto de técnicas e métodos que podem nos auxiliar a fazer de forma automática a coleta dessas informações e estrutura-las de forma que fique mais fácil obter conhecimento, neste trabalho, iremos utilizar a técnica de Análise de Sentimentos, para sabermos o que os usuários da rede social Twitter estão sentindo com relação aos principais candidatos à Presidência da República.

Nesse ano corrente (2014) estamos escolhendo mais uma vez nosso futuro presidente, como é de costume, de quatro em quatro anos voltamos às urnas para escolhermos o futuro líder do nosso país. De acordo com o site do Superior Tribunal Eleitoral[3], em 2014 temos 11 candidatos à chefe da nação, são eles: Aécio Neves da Cunha, Dilma Vana Rousseff, Eduardo Henrique Accioly Campos, Eduardo Jorge Martins Alves Sobrinho, Everaldo Dias Pereira, José Levy Fidelix Da Cruz, José Maria de Almeida, Jose Maria Eymael, Luciana Krebs Genro, Mauro Luís Iasi e Rui Costa Pimenta. Os principais candidatos selecionados para ser fruto de objeto da nossa análise são: Aécio Neves, Dilma Rousseff e Eduardo Campos.

Alguns trabalhos norteiam este trabalho, são eles: Morais e Ambrósio (2007) que traz os conceitos de Mineração de Textos, bem como técnicas e métodos científicos, Barbosa et al (2013) que analisou o sentimento da população online das eleições para presidente dos Estados Unidos em 2012, Sousa (2012) que em seu trabalho aprofunda a análise de opinião na

rede social Twitter e por fim temos o trabalho de Barbosa (2012) que explica como o uso de *hashtags* podem servir como fonte de dados para análise de sentimento.

No primeiro momento, escolhemos um dicionário de *hashtags* para todos os candidatos selecionados para o trabalho. Cada *hashtag* terá um peso de sentimento, positivo, negativo ou neutro. Depois de selecionado o dicionário, iremos para a fase de coleta dos tweets, onde os mesmos serão analisados em seguida. Ao final da análise mostramos os resultados obtidos e o parecer sobre todos os candidatos de acordo com a população *online* do Twitter.

[1] http://noticias.serasaexperian.com.br/facebook-e-lider-entre-redes-sociais-em-fevereiro-no-brasil-de-acordo-com-hitwise/
[2] http://www.techtudo.com.br/noticias/noticia/2014/03/twitter-faz-8-anos-microblog-revela-numeros-sobre-o-brasil-e-o-mundo.html
[3] http://www.tse.jus.br/eleicoes/eleicoes-2014/sistema-de-divulgacao-de-candidaturas

2 TRABALHOS RELACIONADOS

O trabalho aqui apresentado baseia-se em informações, métodos e técnicas coletados de alguns artigos e trabalhos científicos da área de mineração de dados e mineração em texto. Os trabalhos de Morais e Ambrósio (2007), Barbosa et al (2013), Sousa (2012) e Barbosa et al (2012) serviram de norte para o desenvolver da pesquisa. Vejamos a seguir o que cada trabalho trata e como isso contribuiu no trabalho desenvolvido.

Em Morais e Ambrósio (2007) temos o estudo da arte sobre mineração de textos, os conceitos chaves, as técnicas relacionadas e a metodologia de aplicação sobre textos planos, foram bastante utilizados aqui no trabalho apresentado. O processo de descoberta de conhecimento sobre textos apresentado por esse trabalho é dividido em duas etapas, sendo elas: Descoberta de Conhecimento em Dados Estruturados e Descoberta de Conhecimento em Dados Não Estruturados. Entende-se que dados estruturados são todos provenientes de pesquisa, estatísticas e ferramentas, que apresentam a informação de forma organizada. Já os dados não estruturados são todos os dados que não possuem nenhuma uniformidade, geralmente em um volume muito grande, se faz necessário "garimpar" o que vai ser coletado. É nesses dados não estruturados que a mineração de texto trabalha. As técnicas utilizadas pelo autor serviram de base para o trabalho aqui apresentado.

Por sua vez, Sousa (2012) aprofunda o uso da mineração de dados no Twitter, o mesmo objetiva coletar sentimentos dos usuários da rede social com relação a cidade de Campina-SP. Ele utiliza para a extração de dados do Twitter a ferramenta *Twitter Streaming API* e *Twitter Search API*, disponibilizada pela própria rede social. A primeira ferramenta captura os dados em tempo real, enquanto a segunda captura os dados baseados em um período de tempo. As mesmas ferramentas fazem parte do trabalho aqui desenvolvido. Em sua metodologia, o autor tem as seguintes fases: Coleta de Dados, Pré-processamento de Dados, Mineração dos Dados e Análise de Dados, o trabalho em questão também fará uso da mesma metodologia. O que difere o trabalho aqui apresentado com o de Sousa (2012) diz respeito ao uso do algoritmo de classificação, neste trabalho será utilizado o algoritmo de Bayes para a classificação das mensagens e no trabalho de Sousa (2012) o mesmo utiliza o algoritmo SVM, outro ponto que Sousa (2012) difere deste trabalho diz respeito as categorias de sentimento, em seu trabalho o mesmo utiliza somente duas classificações: positivo e negativo, já neste trabalho aqui apresentado, foi escolhida quatro categorias: positivo, negativo, neutro e ambíguo.

Barbosa et al (2013) mostra no seu trabalho a importância da análise de opinião e sentimento na tomada de decisão, baseado no que os usuários do Twitter estão comentando. Com o mesmo viés do trabalho aqui apresentado, Barbosa et al (2013) apresenta uma análise com o uso de *hashtags,* no Twitter, sobre as eleições presidenciais dos Estados Unidos no ano de 2012. O autor coletou uma quantidade significativa de *tweets* que continham a palavra "Obama" em seu texto e em cima deles atribuiu uma classificação baseada no sentimento: Positivo, Negativo, Ambíguo e Neutro. Ao final dessa categorização ele analisa dois aspectos: se a opinião reportada a partir das *hashtags* reflete no sentimento real da população ao longo do tempo e como tem sido a popularidade e propagação dessas hashtags no Twitter.

O trabalho aqui apresentado irá utilizar-se dessa mesma classificação e do processo que foi utilizado por Barbosa et al (2013).

Fechando essa seção de trabalhos relacionados, temos o trabalho de Barbosa et al (2012), que trata de explicar como o uso das *hashtags* podem servir como fonte de dados para análise de sentimento. Neste trabalho o autor tem como principal objetivo verificar se o uso das *hashtags* refletem o sentimento real, assim como temos no trabalho de Barbosa et al (2013). No trabalho o autor definiu um conjunto de hashtags e as categorizou baseados em alguns sentimentos. O mesmo será feito aqui no trabalho presente. Ao final do trabalho foi analisado se o resultado obtido com o uso das redes sociais, refletia na realidade acerca do tema proposto. O mesmo será feito neste trabalho, uma análise de sentimento dos dados do Twitter conflitando com os resultados das eleições presidenciais.

3 FUNDAMENTAÇÃO TEÓRICA

Para um melhor entendimento dos termos e conceitos deste trabalho, a seguir temos uma explanação dos mesmos aqui envolvidos, são eles: Mineração de Dados, com foco em Mineração de Textos, Classificador Naive Bayes e o Twitter e o uso das *hashtags*.

3.1 Mineração de Dados

Segundo Tan, Steinbach e Kumar (2009), mineração de dados pode ser compreendido como o processo no qual é descoberto de forma automática informações úteis em grandes depósitos de dados. Existem diversas técnicas para mineração de dados que são voltadas para agir sobre grande banco de dados com a intenção de descobrir padrões de informações úteis e recentes que podem servir para uma previsão do resultado de uma observação futura, ignorando o uso destas técnicas os dados poderiam ficar nos depósitos de dados sendo não levado em consideração o seu real valor.

Em Berry e Linoff (2004), é apresentado que os algoritmos de mineração de dados foram pensados com fins comerciais. As técnicas de mineração possuem um misto de estatística, ciência da computação e pesquisa em aprendizagem de máquina. As escolhas das técnicas aplicadas na mineração de dados não podem ser engessadas, visto que cada nicho de dados deve ser tratado diferentemente. A mineração de dados é apresentada em duas vertentes, direta ou indireta. A mineração direta busca uma explicação ou categorização sobre um determinado campo de informação, é uma busca por respostas, ou informações que leve as mesmas. Já a vertente indireta, busca por padrões, similaridades entre os dados, visando prever resultados futuros baseados nesses padrões.

3.2 Mineração de Textos

A mineração de dados baseada em texto, ou simplesmente, mineração de textos pode ser entendida como uma busca por informações significativas a partir de uma grande quantidade de textos escritos em linguagem natural. Assim, como na mineração de dados convencional, temos a aplicação de algoritmos e métodos de aprendizagem de máquina, bem como o uso de estatísticas aplicadas aos textos, visando buscar informações úteis. Com esse propósito, é preciso pré-processar os textos para que os mesmos tenham uma certa coesão (HOTHO, 2005).

Para Fayyad, Piatetsky-Shapiro e Smyth (1997) os textos apresentam, na sua maioria, uma maior dificuldade na extração de informações se comparados aos dados organizados em banco de dados. Isso se deve ao fato de que no banco de dados os dados geralmente estão organizados e estruturados em tabelas e as mesmas possuem algumas relações, o que nos garante mais coesão nas informações.

Diversos autores discorrem sobre as etapas de mineração de textos, alguns propõem duas, três ou até mais etapas distintas, porém aqui no trabalho em questão utilizamos a abordagem de Aranha (2013). Essa abordagem divide o processo de mineração de dados em 5 etapas, sendo elas: Coleta, Pré-Processamento, Indexação, Mineração de Dados e Análise dos Resultados.

Aranha (2013) discorre sobre as etapas da seguinte maneira: Coleta é a etapa onde é constituída a fonte de dados que serão utilizada na mineração; Na etapa de Pré-Processamento trabalham-se os dados para o processo computacional, remover os dados irrelevantes é um exemplo de técnica aplicada nessa fase; Na terceira etapa, Indexação, o foco é categorizar e separar os dados visando facilitar a busca e acesso, no trabalho aqui apresentado dividimos a coleta dos três candidatos, Aécio, Dilma e Marina, e em cada candidato foram novamente separados pelos meses; A quarta etapa, Mineração de Dados, é onde ocorre a mineração de dados em si, através de técnicas de cálculos e inferências; A quinta e última etapa desse processo, Análise dos Resultados, conta com a interferência e entendimento humano sobre os resultados obtidos. Vide figura abaixo para visualizar o processo.

Figura 1 - Etapas do processo de Mineração de Textos

Fonte: Aranha (2013)

3.3 Teorema de Bayes e Classificador Naive Bayes

Até o começo do século XVIII a grande maioria dos problemas probabilísticos relacionados a eventos com condições específicas estavam bem resolvidos. Porém começaram a surgir novos problemas mais complexos que utilizavam os resultados dos problemas anteriores como entrada para novos problemas, novos questionamentos. Partindo desse ponto, Thomas Bayes, um ministro inglês do século XVIII, começou a pensar sobre e formulou um pensamento até então revolucionário para a época (OGURI, 2006).

Para exemplificar o Teorema de Bayes, Tan, Steinbach e Kumar (2009, p. 270-271) apresentam em sua obra da seguinte maneira:

> *Considere um jogo de futebol entre duas equipes rivais: a Equipe 0 e a Equipe 1. Suponha que a Equipe 0 vença 65% do tempo e a Equipe 1 os jogos restantes. Entre os jogos vencidos pela Equipe 0, apenas 30% deles vêm de jogos no campo da Equipe 1. Por outro lado, 75% das vitórias da Equipe 1 são obtidas jogando em casa. Se a Equipe 1 receber o próximo jogo entre as duas equipes, qual equipe provavelmente vencerá?*
>
> Esta questão pode ser respondida usando-se o bem conhecido teorema de Bayes. Para que nada fique faltando, começamos com algumas definições básicas da teoria da probabilidade. Os leitores que não estiverem familiarizados com conceitos de probabilidade podem ver no Apêndice C uma breve revisão deste tópico.
>
> Suponha que *X* e *Y* sejam um par de variáveis aleatórias. Sua probabilidade junta, $P(X = x, Y = y)$, se refere à probabilidade da variável *X* receber o valor *x* e a variável *Y* receber o valor de *y*. Uma probabilidade condicional é a de que uma variável aleatória receba um determinado valor dado o resultado de outra variável aleatória seja conhecido. Por exemplo, a probabilidade condicional $P(Y = y, X = x)$ se refere à probabilidade da variável *Y* receber o valor *y*, dado que a variável *X* tenha o valor *x*. As probabilidades condicionais e juntas estão relacionadas da seguinte forma:
>
> $$P(X, Y) = P(P\backslash X) \, x \, P(X) = P(X\backslash Y) \, x \, P(Y).$$
>
> Reorganiza as duas últimas expressões na Equação 5.9 leva à seguinte fórmula, conhecida como teorema de Bayes:
>
> $$P(Y|X) = \frac{P(X|Y)}{P(X)}.$$
>
> O teorema de Bayes pode ser usado para resolver o problema da previsão declarado no início desta seção. Por conveniência de notação, suponhamos que X seja a variável aleatória que represente a equipe local da partida e que Y seja a variável aleatória que representa o vencedor da partida. Tanto x quanto Y podem receber valores do conjunto {0, 1}. Podemos resumir as informações dadas no problema da seguinte maneira:
>
> A probabilidade da Equipe 0 vencer é $P(Y = 0) = 0{,}65.$
>
> A probabilidade da Equipe 1 vencer é $P(Y = 1) = 1 - P(Y = 0) = 0{,}35.$
>
> A probabilidade da Equipe 1 ser a local e vencer é $P(X = 1|Y = 1) = 0{,}75.$
>
> A probabilidade da Equipe 1 ser a local e a vencedora ser a Equipe 0 é $P(X = 1|Y = 0) = 0{,}3.$
>
> Nosso objetivo é calcular $P(Y = 1|X = 1)$, que é a probabilidade condicional da Equipe 1 vencer a próxima partida que jogará em casa e comparar isso a $P(Y = 0|X = 1)$. Usando o teorema de Bayes, obtemos

$$P(Y=1|X=1) = \frac{P(X=1|Y=1)\times P(Y=1)}{P(X=1)} = \frac{P(X=1|Y=1)\times P(Y=1)}{P(X=1,\ Y=1)+P(X=1,Y=0)} = \frac{P(X=1|Y=1)\times P(Y=1)}{P(X=1|Y=1)\,P(Y=1)+P(X=1|Y=0)\,P(Y=0)} = \frac{0.75\times 0.35}{0.75\times 0.35+\ 0.3\times 0.65} = 0.5738,$$

onde a lei da probabilidade total (veja a Equação C.5 na página 722) foi aplicada na segunda linha. Além disso, $P(Y=0,X=1)=1-P(Y=1|X=1)=0,4262$. Já que $P(Y=1,X=1)>P(Y=0|X=1)$, Equipe 1 possui mais chance do que a Equipe 0 de vencer a próxima partida.

Para Oguri (2006) o Classificador Naive Bayes, utilizado para classificar dados baseados em um modelo computacional, é um dos mais utilizados no mundo para o aprendizado de máquinas (Learning Machines). O classificador é tido como Naive (ingênuo) pois assume que a informação de um determinado evento, não serve de informação para outro evento, dentro do mesmo contexto. Neste trabalho iremos utilizar o modelo chamado Modelo Multinomial, onde assumimos que cada documento é representado por um vetor de atributos inteiros caracterizando o número de vezes que cada especificidade ocorre no documento.

Para facilitar o entendimento, imagine uma frase, com todas suas palavras e expressões. Ao gerar o documento o algoritmo irá percorrer toda a frase contando as palavras e as colocando em um vetor, com seu conteúdo e suas ocorrências.

Tan, Steinbach e Kumar (2009), a classificação utilizando as técnicas de Bayes consiste em fazer um registro de teste que mais tarde será aplicado a um conjunto de dados para que sejam classificados baseados no que o modelo gerado na fase de teste aprendeu.

3.4 Twitter

Russel (2013) descreve o Twitter como um serviço web de microblog, disponível gratuitamente e onde a pessoa pode expressar suas ideias e sentimentos com apenas 140 caracteres por mensagem. Sua proposta de divulgação de mensagem simples e rápida tornou-a uma ferramenta mais utilizada no mundo, com mais de 500 milhões de usuários e mais de 100 milhões de usuários que usam diariamente. O Twitter mostra um conceito bastante interessante no que se refere a sua taxonomia, separação dos dados em categorias, o uso da *hashtag*. A mesma é conceituada como uma palavra-chave para a mensagem enviada, deve ser prefixada com o uso do caractere # (cerquilha), exemplo disso seria a hashtag #SouBrasileiro. Seu uso não é obrigatório. Dentro do Twitter ele cria um link na(s) hashtag(s) da mensagem com o intuito de separar as diferentes categorias.

O uso das *hashtags* são fundamentais no trabalho aqui apresentado é através delas que podemos separar as diferentes categorias e sentimentos. Outra funcionalidade que chama atenção no Twitter é o recurso de Retweet, onde um usuário pode se utilizar de uma mensagem outro usuário e retransmiti-la em sua linha do tempo. A mensagem retransmitida é prefixada com “RT” seguido do nome do usuário original (@usuário). Esse recurso é válido neste trabalho, visto que o usuário pode expressar sua opinião e pensamentos através de uma mensagem de outro, ou seja, ele compartilha do mesmo sentimento do autor da mensagem original.

4 PROCEDIMENTOS

Os procedimentos aqui adotados nesse trabalho foram divididos em etapas distintas para facilitar o entendimento. São eles:

- Escolha dos candidatos;
- Descoberta de *hashtags*;
- Coleta de *tweets*;
- Pré-processamento dos *tweets* coletados;
- Análise dos Sentimentos;
- Avaliação dos resultados obtidos.

4.1 Escolha dos Candidatos

A ideia é escolher os candidatos mais bem classificados por pesquisa popular, antes e durante as eleições. Foi estipulado que somente seriam analisados os três primeiros colocados para concorrer ao cargo de presidente da república. De acordo com pesquisa realizada entre maio de 2014 e julho do mesmo ano pelo Instituto de Pesquisa Datafolha, os principais candidatos são: Dilma Rousseff (PT), Aécio Neves (PSDB) e Eduardo Campos (PSB). Esses foram os candidatos escolhidos para começar o trabalho, porém, fatidicamente no dia 13 de agosto de 2014 o candidato Eduardo Campos sofreu um acidente de avião e veio a falecer, em seu lugar como candidato pelo partido PSB, foi lançada a candidatura de Marina Silva, logo foi alterado o objeto da pesquisa.

Sendo assim, tivemos uma alteração nos objetos desse trabalho, agora sendo selecionados: Dilma Rousseff (PT), Aécio Neves (PSDB) e Marina Silva (PSB).

4.2 Descoberta de *hashtags*

Foram realizadas uma séria de coletas prévia de tweets relacionados para cada candidato selecionado como objeto da pesquisa. Foram utilizadas nessa coleta os nomes dos candidatos e as siglas dos partidos. Foram obtidos uma massa expressiva de dados, felizmente. Depois dessa etapa de coleta foi realizado um processo de descobrir quais as hashtags foram utilizadas em cada tweet e a ocorrência das mesmas na coleta geral, chegando assim a formar um dicionário de hashtags para cada candidato. Esse dicionário será utilizado na etapa de coleta de tweets com o intuito de filtrar a massa de dados, agregando mais valor a coleta.

4.3 Coleta de tweets

Baseado no dicionário gerado pela etapa anterior, iniciou-se a coleta dos tweets com o uso da API Streaming disponibilizada pelo Twitter. A linguagem utilizada pelo programa é Ruby, por se tratar de uma linguagem de alto nível e de fácil implementação. O programa procura tweets em tempo real baseados nas hashtags do dicionário de cada candidato. Os dados coletados foram o identificador do tweet (id) e o conteúdo do tweet em si (mensagem).

As coletas são separadas em arquivos para cada candidato no formato .tsv (Tab Separated Values), que separa por tabulação os dados da captura. As coletas iniciam-se durante todo o processo eleitoral e estendem-se até uma semana após o fim do primeiro turno das eleições.

4.4 Pré-processamento dos *tweets* coletados

Após a coleta do tweets de cada candidato os dados serão necessários eliminar alguns tweets que não fazem parte da pesquisa, como links, propagandas e tweets em outras línguas, como espanhol e inglês, que podem confundir o classificador.

Essa fase visa formatar os dados para que sejam utilizados no classificador, é necessário deixar os dados mais limpos e expressivos. Ao final dessa fase os dados estão mais simples, porém com o valor real do que queremos analisar. Esses serão os dados no qual o classificador irá trabalhar sobre.

4.5 Análise dos Sentimentos

Definidos quais tweets serão utilizados é necessário fazer um dicionário que classifica o conteúdo dos tweets em positivos, negativos, neutros ou ambíguos. Nessa fase foi utilizado o algoritmo de Naive Bayes, que nada mais é do que um classificador de probabilidades baseado na aplicação do teorema de Bayes com forte independência de suposições. Vale salientar que o algoritmo em questão é um dos mais aplicados ao aprendizado de máquina. O algoritmo foi escolhido por seu nível de confiança sobre os resultados obtidos e pela sua facilidade de aplicação com textos.

Serão utilizadas duas ferramentas para auxiliar o trabalho de mineração de dados do trabalho aqui apresentado, são elas: Hadoop e Mahout. O Hadoop é uma implementação de código aberto do paradigma de programação Map-Reduce. Map-Reduce é um paradigma de programação introduzido pelo Google para processar e analisar grandes conjuntos de dados. Todos esses programas que são desenvolvidos nesse paradigma realizam o processamento

paralelo de conjuntos de dados e podem, portanto, ser executados em servidores sem muito esforço. A razão para a escalabilidade desse paradigma é a natureza intrinsecamente distribuída do funcionamento da solução. Uma grande tarefa é dividida em várias tarefas pequenas que são então executadas em paralelo em máquinas diferentes e então combinadas para chegar à solução da tarefa maior que deu início a tudo. Os exemplos de uso do Hadoop são analisar padrões de usuários em sites de e-commerce e sugerir novos produtos que eles possam comprar. Já o Mahout, ou Apache Mahout, é um projeto de código fonte aberto com o objetivo primário de utilizar algoritmos de aprendizagem por máquina escaláveis. O Mahout contém implementações para armazenamento em cluster, categorização, CF, e programação evolucionária. Além disso, quando prudente, ele usa a biblioteca do Hadoop para permitir que o Mahout escale de forma efetiva na nuvem o que possibilita um processamento de grandes quantidades de dados.

Os passos a seguir foram os mesmos utilizados por Ngoc (2013) em seu tutorial. Dito isto temos os seguintes passos para a utilização do algoritmo:

- Transformar o arquivo de captura para sequência;
- Enviar os arquivos para o servidor do Hadoop;
- Transformar a sequência em vetores;
- Separação dos dados de treino e teste;
- Treinar o computador;
- Testar e validar o modelo.

Iremos transformar os nossos arquivos de capturas em arquivos de sequência, pois o Mahout não conhece a forma como nossos dados foram capturados. Em seguida, iremos enviar os dados, agora em sequência para o servidor de Hadoop afim de que o mesmo gere os modelos e gere as classificações. Utilizando o Mahout juntamente com o Hadoop iremos transformar os arquivos de sequências em vetores, estes vetores são os mais diversos, desde a contagem de termos, frequência com as quais eles aparecem no texto até a lista de todos os termos utilizados. Gerado os vetores, iremos separar os mesmos para treinar o computador e depois testar para averiguar se o mesmo está funcionando a contento. Nessa fase de separação informamos qual a porcentagem será teste e treino, no caso foi utilizada 70% do conteúdo para treino e 30% para teste, ambos valores foram definidos por serem valores bastante utilizados em outras pesquisas da área. Treinamos o computador, com essa massa de dados anteriormente citada, ao final do processo o mesmo gera um modelo de classificação que

poderá ser utilizado em novas classificações. Depois de treinado e gerado o modelo, utilizaremos os mesmos agora para testar a outra massa de dados que foi separada.

No caso do trabalho aqui em questão, foi utilizada uma massa de 3000 entradas para treino e teste, e mais 3000 para uma verificação posterior. Vale lembrar que o processo aqui citado é realizado para cada candidato, pois cada um dos mesmos tem seu modelo próprio de classificação.

4.6 Avaliação dos resultados obtidos

Com os dados das avaliações geradas pelo classificador gerados na fase anterior, que expressa o que os usuários online sentem, iremos confrontar com os fatos reais que estão ocorrendo nas eleições. Como exemplo podemos verificar se o que as pessoas do Twitter falam de algum candidato correspondem as pesquisas e atos que acontecem no mundo real.

Após as eleições veremos se os sentimentos com relação ao candidato escolhidos correspondem aos mesmo dos usuários do Twitter. E qual o sentimento para cada candidato não eleito.

5 DESENVOLVIMENTO/RESULTADOS

5.1 A Coleta

O trabalho em questão coletou tweets de todos os três candidatos no período de agosto, setembro e início de outubro (05 de Outubro), ou seja, boa parte do período do processo eleitoral no primeiro turno. As coletas se deram de agosto em diante, pois antes disso havia sido coletado do candidato Eduardo Campos que por infelicidade do destino veio a falecer, como dito anteriormente. As coletas dos tweets visavam capturar mensagens que contivessem as seguintes *hashtags*:

Para a candidata Dilma Roussef: #DilmaDeNovo, #EuSouDilma, #DilmaSuperSimples, #ForaDilma, #PresidentaCatifunda, #DilmaDerrotada, #EncontroComDilma, #ForaDilmaLulaPT, #SouMaisDilma, #Brasil13, #DilmaDoChefe, #DilmaPresidenteDoFracasso, #DilmaRejeitada, #DiaDeVaiarADilma, #DilmaMudaMais, #EstudantescomDilma, #foradilma, #Dilmais, #VaiTerDilma, #DilmaCorrupta, #BomDilma, #PTpuraMentira, #dilma2014, #Dilmalice, #ForaPT2014, #DeixaAmeninaDilma, #Dilmafujona, #DilmaArregona, #DilMentirosa, #TVComuna, #Dilmais4, #DILMAFUDIDA, #VotoDilmaPorque, #DilmanoRadio, #culpadaDilma, #DilmaFica, #Dilma13MaisEmprego, #Dilma13Neles, #ONordesteeDilma, #dilmanatv, #dilma13maisfuturo, #efeitodilma, #DimaDeNovo, #DilmaQuebrouoBrasil, #Dilma13Presidenta, #DilmaPareceUmDiaboLoiro, #LulaeDilma13Neles, #VOTODILMAIS, #dilmanao, #dilmasim, #ficaDilma, #dilma13, #soudilma13, #dilmapresidente, #eusoudilma13, #foradilma;

Para a candidata Marina Silva: #SouMarina40, #euemarina40, #ForaMarina, #SOUMARINA40, #Marina40, #MarinaPresidente40, #MarinaPresidenta, #soumarina40, #MalafaiaMandaMarinaObedece, #AgendaMarina, #MarinaSilva, #marinasilva40, #SomosMarina40, #MarinaNoJN, #EuNaoVouDeMarina, #votomarina, #vaiMarina, #MarinaVoltaAtras, #Desafio40, #MarinaSilvaIndecisa, #Marinasaifora, #SomosTodos40, #FecheiComMarina, #Presidente40, #marinanao, #marinasim, #marina40, #soumarina40, #marinapresidente, #naovamosdesistir, #naovamosdesistirdobrasil, #MarinaSilvaPresidenta, #foramarina;

E para o candidato Aécio Neves: #AECIOPORTO", "#SouAecio", "#AecioPresidente", "#Aecioportos", "#EstamosComAecio", "#Aecio45", "#ForaAecio", "#EquipeAN", "#aeciopresidente", "#QuemVotaTassoVotaAecio", "#AecioManiaNacional",

"#somosaecio", "#HELICOCA", "#Aecioneves", "#AecioMudaBrasil", "#ESTAMOS_COM_AECIO", "#AecioNever", "#aecio45", "#AecioNaGlobo", "#FechadoComAecio", "#DesconstruindoAecioNoJN", "#aecionojn", "#EuVoto45", "#AECIOporto", "#EstoucomAecio", "#SomosAecio", "#chupaaecio", "#AecioTecoTeco", "#aeciofacts", "#VcAecio", "#vote45", "#AecioNaoResponde", "#tamojuntosaecio", "#aeciotipotitanic", "#Sou45", "#EuVotoNoAecioNeves", "#45Aecio", "#vaiaecio", "#EuVotoAecio45", "#SomosMaisAecio", "#SouAecio45", "#EuSouAecio45", "#AecioNevespresidente", "#SouMaisAecio", "#aeciooumarina", "#AecioTaEmMinas", "#AecioNoJornalDaGlobo", "#SomosAecio45", "#ElejaAecio", "#FechadocomAecio", "#semprecomaecio", "#somosaecio45", "#AecioNoEstadao", "#vamosAecio", "#calaabocaaecioneves", "#aecionao", "#aeciosim", "#aecionever", "#psdbNuncaMais", "#aeciopresidente", "#euvoudeaecio", "#e45", "#mudabrasil", "#aecio45", "#souaecio45", "#AecioCheiraADerrota", "#foraaecio".

O tamanho de dados (em megabyte) obtido para candidato foi de: 20 megabytes para a candidata Dilma Roussef, 32 megabytes para Marina Silva, e 40 megabytes para o candidato Aécio Neves.

Conforme tabela abaixo podemos ver a quantidade de dados coletados em cada período, baseado nos filtros de hashtag acima descritos:

Tabela 1 - Resultado da Coleta de Dados

Candidato	**Agosto**	**Setembro**	**Outubro**	**Total**
Dilma Roussef	10692	21904	3299	**35895**
Marina Silva	12051	45254	3969	**61274**
Aécio Neves	11348	50174	9399	**70921**

Analisando a coleta é perceptível primeiramente que em no mês de agosto os candidatos se mantinham com um número semelhante de mensagens referentes a suas hashtags. No mês seguinte, setembro os candidatos Marina Silva e Aécio Neves tiveram um número bastante expressivo de mensagens, com mais que o dobro da candidata Dilma Roussef. Em outubro, período esse de menos de uma semana, tivemos uma discrepância na quantidade do candidato Aécio Neves, com quase três vezes o volume de mensagens dos demais candidatos.

5.2 Validação do Modelo

Foram separados 3000 tweets para realização do processo de geração do modelo, modelo este que é utilizado para a classificação dos demais dados. O treino do modelo seguiu 4 critérios: Positivo, quando o sentimento contido na mensagem era de apoio ao candidato em

questão; Negativo, quando o sentimento contido na mensagem era contra o candidato em questão ou quando enaltecia outro candidato senão o que estava em questão; Neutro, quando o conteúdo da mensagem era uma ação da campanha, ou links informativos sobre o candidato; Ambíguo, quando não era possível auferir com certeza o sentimento da mensagem, podendo variar entre positivo e negativo, dependendo da interpretação do leitor.

Após a fase de separação de teste e treino, onde foram separados 70% destes tweets para treino e 30% para teste, chegamos a um coeficiente de precisão do modelo, conforme vemos na tabela abaixo, este coeficiente, dado em porcentagem, nos mostra que para uma determinada quantidade de entradas o modelo classifica corretamente X% dos dados.

Tabela 2 - Coeficientes de Precisão do modelo para cada candidato

Candidato	**Coeficiente de Precisão do Modelo**
Dilma Roussef	83%
Marina Silva	79%
Aécio Neves	87%

Para verificar mais afinco a precisão, foi realizado um teste manual com o modelo de cada candidato, visando verificar se o modelo automático está classificando corretamente as entradas que lhe são fornecidas. O teste foi realizado separando 100 tweets aleatórios de cada candidato e utilizando o classificador automático foi solicitado que o mesmo classificasse as entradas baseadas no modelo já estabelecido. Ao final deste processo automático foi realizada uma conferência manual dessa classificação e temos como resultado o quadro abaixo, onde mostra o confrontamento do coeficiente de precisão do modelo e a taxa de acertos, verificada manualmente, do mesmo sobre as entradas utilizadas.

Tabela 3 - Comparação entre coeficientes de precisão e taxa de acertos

Candidato	**Coeficiente de Precisão do Modelo**	**Taxa de Acertos**
Dilma Roussef	83%	80%
Marina Silva	79%	82%
Aécio Neves	87%	90%

Correlacionando o coeficiente de precisão e a taxa de acertos, vimos que os modelos gerados aproximaram muito do esperado pelo coeficiente de precisão, o que nos garante dizer que nosso classificador se encontra em condições de auferir de forma satisfatória qualquer quantidade de dados que sejam utilizadas na entrada.

5.3 As palavras mais utilizadas (Top Words)

Baseado no modelo de classificação descrito na etapa anterior, retiramos as palavras mais utilizadas em cada categoria e no geral. A tabela abaixo nos mostra o resultado para cada candidato.

Tabela 4 - Palavras mais utilizadas para cada candidato

Candidato	Geral	Positivo	Negativo	Ambíguo	Neutro
Dilma Roussef	http, foradilma, forapt, rt, dilma	dilmadenovo, http, dilma13maisfuturo, dilmamudamais, dilma13	foradilma, forapt, rt, http, não	Dilma13, lucianagenro, eleições2014, aecio45, marina40	Brasil13, lulanews, marina, http, dilma
Marina Silva	marina, marina40, soumarina40, http, rt	soumarina40, http, rt, marina, marina40	foramarina, foradilma, nemmarinanemdilma, marinacensura, aeciodevirada	alckmin, eleitores, amigo, bancos, pt	silva_marina, rt, http, 40presidente, marina
Aécio Neves	equipean, http, rt, aecioneves, aeciopresidente	aeciopresidente, rt, http, aecio45, brasil	aecionever, matarjamais, fhc, desastrosa, http	mudabrasil, marina40, seguindo, dilma13denovo, ofertas	aecioneves, rt, http, aécio, equipean

É perceptível que em todas as principais palavras gerais vemos duas que se repetem em todos os candidatos, são elas as palavras/expressões “rt” e “http”, o que pode nos deixar claro duas coisas: a palavra/expressão “rt” é uma expressão utilizada na rede social Twitter para designar quando uma pessoa repassa uma mensagem de outra pessoa, é a sigla para a expressão Retweet, ou seja temos um grande volume de tweets que são mensagens repassadas por outros usuários que sentem o mesmo que o autor original quis passar; a palavra/expressão “http” nos remete ao uso de links para informações externas a rede social em questão, geralmente são utilizadas para disseminar notícias ou ações publicitárias de cada candidato.

5.4 Nuvem de Palavras (Word Cloud)

Baseado nos dados coletados de cada candidato foram geradas as nuvens de palavras para cada um destes. As nuvens de palavras nos ajudam a entender as principais palavras/expressões que são mais utilizadas nas mensagens coletadas. Quanto maior for o tamanho da palavra, mais vezes esta ocorreu nos dados coletados. A seguir poderemos ver as nuvens de cada candidato de uma forma geral.

Vejamos as nuvens de palavras na seguinte ordem: Dilma Roussef, Marina Silva e por fim Aécio Neves.

Figura 2 - Nuvem de Palavras - Dilma Roussef

Fonte: Dados da pesquisa

As palavras/expressões que mais chamam a atenção na nuvem de palavra da candidata Dilma Roussef foram: “ForaDilma” e “ForaPT”, o que nos mostra um descontentamento com a candidata por parte dos usuários. Fica evidente que a candidata foi alvo de críticas e não aceitação por parte das pessoas que enviaram suas mensagens. Em segundo plano temos um outro conjunto de palavras que expressam apoio e aversão a candidata, ambos conjuntos do mesmo tamanho.

Figura 3 - Nuvem de Palavras - Marina Silva

Fonte: Dados da pesquisa

Analisando a nuvem de palavras da candidata Marina Silva podemos destacar a heterogeneidade das palavras e não existem palavras/expressões que se destacam de forma

ímpar, mas temos um conjunto destas: “ForaMarina”, “Marina40”, “MarinaSilva”, “MarinaDeVerdade”, “MarinaCensura” e “MarinaPresidente”. Este conjunto mostra sentimentos bastante diferentes, alguns remetem ao apoio a candidata e outras não. Fica evidenciado que não existe uma constância com relação ao sentimento dos usuários com relação a candidata, diferente da candidata anterior.

Figura 4 - Nuvem de Palavras - Aécio Neves

Fonte: Dados da pesquisa

Por fim no candidato Aécio Neves temos como palavra/expressão principal: “EquipeAN” e um conjunto de outras que mostram uma incidência não tão alta, mas que merecem destaque, são elas: “MudaBrasil”, “Aécio45”, “AécioPraMudar”, “EuVotoAécio45”, “SouAécio45” e “Sou45”. Analisando a palavra/expressão “EquipeAN” notamos que as mensagens eram postadas por pessoas que se identificavam com as propostas do candidato e se viam como uma equipe, muito embora muito de suas mensagens remetam a links meramente informativos. O conjunto subsequente destacado mostra que o sentimento por sua vez destes é positivo, contrário ao que foi observado nos outros candidatos.

5.5 Classificação e Análise dos Sentimentos

Depois de gerado todos esses dados preliminares o trabalho aplicou o modelo gerado em etapas anteriores à classificação da massa de dados de cada candidato de acordo com o período, é preciso salientar aqui que os modelos têm um certo grau de precisão, porém os valores auferidos podem não corresponder em sua totalidade ao sentimento real proposto pelos autores das mensagens, mas é garantido que sua grande maioria é classificada corretamente, conforme visto na seção de validação do modelo.

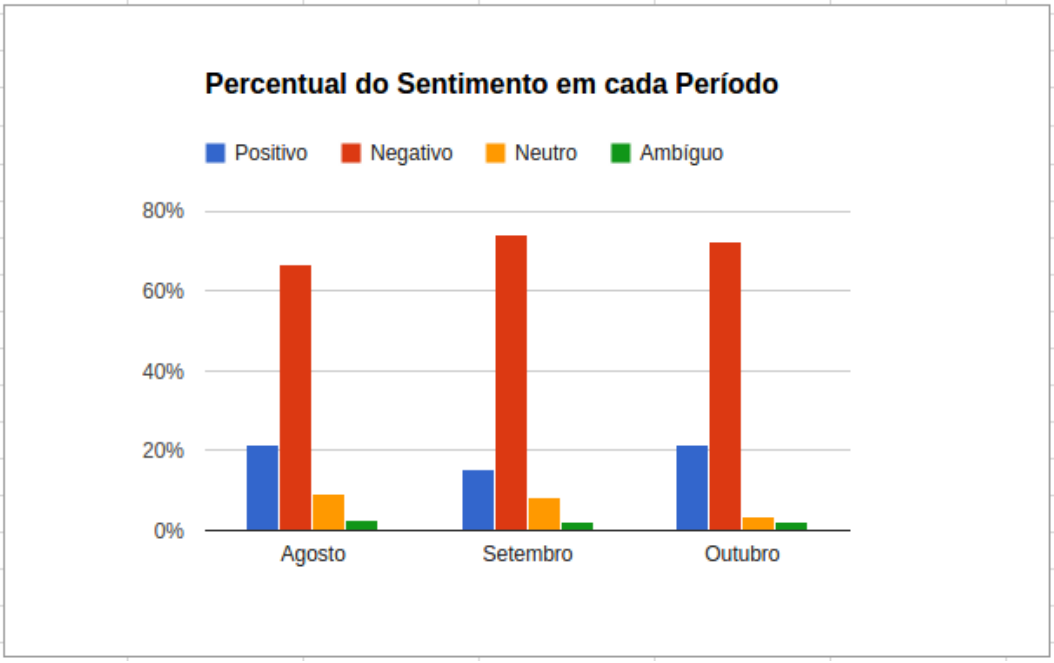
Abaixo temos a tabela com os valores percentuais de cada candidato em cada período, seguidos dos respectivos gráficos e análise sobre estes.

5.5.1 Dilma Rousseff

Tabela 5 - Sentimentos ao longo dos períodos (percentuais) - Dilma Roussef

	Positivo	Negativo	Neutro	Ambíguo	Total
Agosto	21,40151515	66,75189394	9,166666667	2,679924242	100
Setembro	15,28384279	74,00597564	8,421052632	2,289128936	100
Outubro	21,46978439	72,57819617	3,704828424	2,247191011	100

Figura 5 - Gráfico de Sentimento em cada período - Dilma Roussef



Fonte: Dados da pesquisa

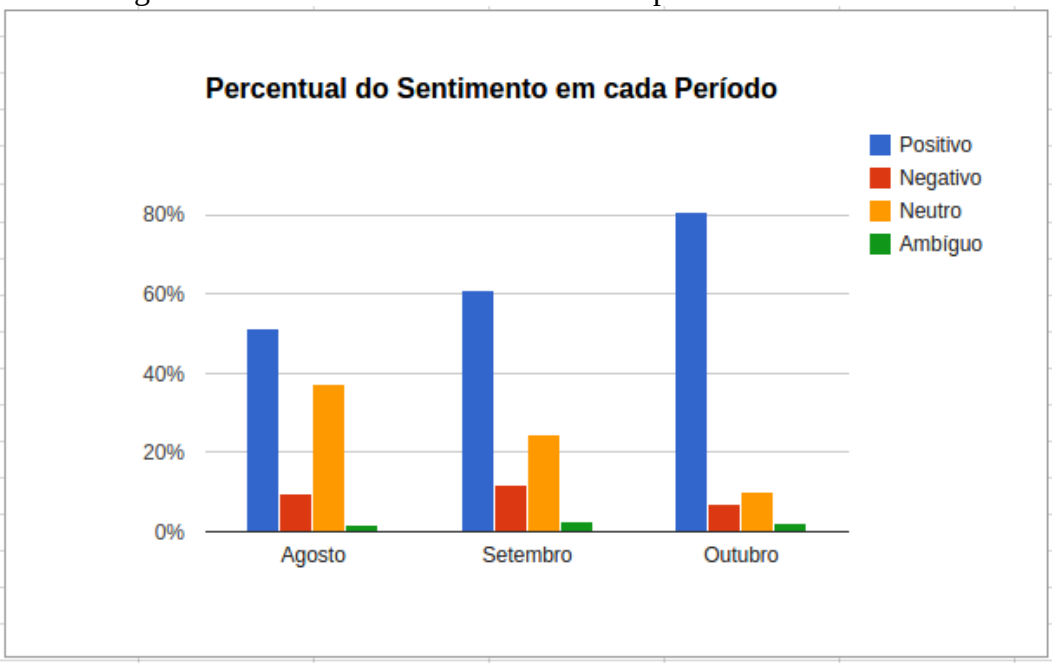
É notório que a candidata a presidente Dilma Roussef não possui grande aceitação e isto fica evidente analisando o gráfico acima, em todos os períodos a maioria das mensagens (mais de 60% na média dos períodos) das mensagens postadas são de cunho negativo para com a candidata. O que contrasta com as pesquisas dos principais órgãos de pesquisa, que apontavam uma vitória de Dilma Roussef ainda no primeiro turno. O parâmetro positivo mantém a média de 20% no total, o que indica que somente uma pouca parcela das mensagens foram de apoio, fato esse que pôde ser observado analisando a nuvem de palavras da mesma. Esse cenário contrasta com o resultado das eleições no primeiro e no segundo turno onde a candidata saiu vitoriosa, o que nos chama a atenção para fazer uma análise maior do perfil do usuário da rede social em questão, no caso o Twitter.

5.5.2 Marina Silva

Tabela 6 - Sentimentos ao longo dos períodos (percentuais) - Marina Silva

	Positivo	Negativo	Neutro	Ambíguo	Total
Agosto	51,10428429	9,805712388	37,48754567	1,602457655	100
Setembro	61,09359737	11,71077169	24,70029305	2,495337892	100
Outubro	80,53521838	6,99318354	10,17419843	2,297399647	100

Figura 6 - Gráfico de Sentimento em cada período - Marina Silva



Fonte: Dados da pesquisa

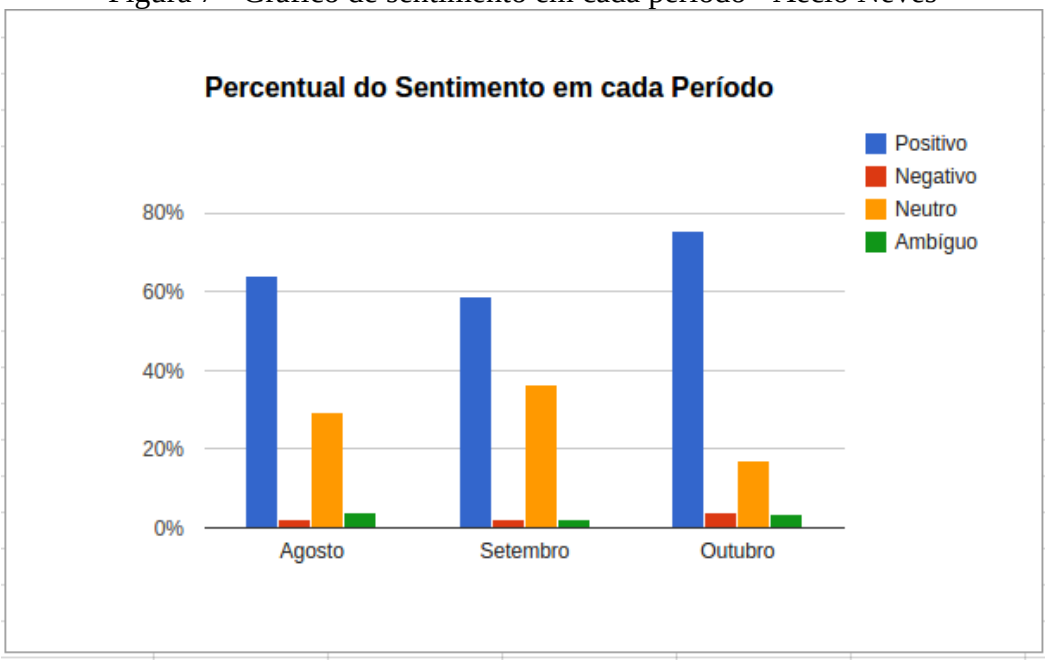
Marina Silva por sua vez apresenta uma crescente no que se refere a mensagens positivas, mesmo que sua nuvem de palavras não demonstre isso. É visível que a mesma aumentou consideravelmente o valor positivo das mensagens de cunho positivo. Outro fato que chama atenção ao analisar o gráfico foi que a taxa de rejeição, mensagens negativas, mantiveram-se em um nível muito baixo. Analisando as mensagens de cunho neutro, aquelas que não expressam nenhum sentimento, decresceu muito no último período. Vale salientar que Marina Silva assumiu a candidatura depois que o seu aliado político, Eduardo Campos, faleceu. Ou seja, no período de agosto, as maiorias das mensagens vinculadas à Marina eram de links de notícia, o que pode ser verificado na seção de palavras mais utilizadas, descrita anteriormente. É possível observar que enquanto a taxa de mensagens de cunho neutro cai a de cunho positivo aumenta, o que nos mostra que as ações de divulgação estão fazendo efeito positivo. Mas não foi o suficiente para a candidata ir para o segundo turno das eleições.

5.5.3 Aécio Neves

Tabela 7 - Sentimentos ao longo dos períodos (percentuais) - Aécio Neves

	Positivo	Negativo	Neutro	Ambíguo	Total
Agosto	63,87836051	2,379903041	29,59012781	4,151608638	100
Setembro	58,81321135	2,393908589	36,47072894	2,322151129	100
Outubro	75,44476404	3,952274422	16,94897198	3,65398956	100

Figura 7 - Gráfico de sentimento em cada período - Aécio Neves



Fonte: Dados da pesquisa

Por fim temos o candidato Aécio Neves, este por sua vez apresenta um alto nível de mensagens positivas, que se mantiveram sempre em sua maioria em taxas superiores a 60%, detalhe para a baixa de setembro, onde o candidato nas pesquisas aparecia em 3º colocado. O critério neutro permeia de forma incisiva o gráfico, o que mostra que muitas mensagens relacionadas ao candidato eram de cunho informativo (notícias, informações de campanha). Note que no período de setembro houve um aumento deste critério (neutro) o que pode caracterizar um aumento de informações e ações de campanha do candidato. A sua taxa de rejeição, mensagens do critério negativo, se mantiveram sempre abaixo dos 5%, o que mostra que a maioria das mensagens não rejeitavam o candidato em questão. No último período é importante notar que enquanto o critério neutro tem uma baixa considerável, o critério positivo chega ao seu máximo, o que nos leva a concluir que as ações de divulgação e de campanha do candidato surtiram efeito. Isto fica evidente quando sai o resultado da eleição no primeiro turno onde o candidato conseguiu sair em segundo lugar e disputou o segundo turno em uma eleição bastante acirrada.

6 CONCLUSÃO E TRABALHOS FUTUROS

O problema deste trabalho era tentar auferir qual o sentimento por trás das mensagens postadas na rede social Twitter sobre os candidatos a presidente da república do Brasil no ano de 2014 no primeiro turno, foram utilizadas técnicas de mineração e classificação de dados. O trabalho coletou, separou, tratou os dados e os classificou para candidato a fim de investigar o que a população online, da rede social em questão, sentia acerca de cada candidato.

O resultado foi bastante satisfatório, tivemos uma grande perda durante o processo, a morte do candidato Eduardo Campos, o que mudou um pouco o rumo do trabalho, já que em seu lugar a candidata Marina Silva assumiu a campanha e tivemos que alterar um dos candidatos no meio do processo.

Abaixo temos um resumo do sentimento analisado de cada candidato baseado na pesquisa aqui realizada.

Sobre a candidata, e atual presidente da república, Dilma Roussef podemos destacar, baseado nos dados, que a mesma possuía um alto índice de rejeição por parte dos usuários da rede social em questão, Twitter, visto a classificação das mensagens, visto a nuvem de palavras e as palavras mais utilizadas (top words). A candidata desde o primeiro período (agosto) já apresentava um número muito alto de mensagens negativas referente a ela, o que se seguiu até o último período, com um aumento no segundo período (setembro)e uma leve que queda no terceiro período (outubro).

Seguindo a ordem, temos Marina Silva, que assumiu a campanha do seu falecido parceiro político, Eduardo Campos, faltando pouco mais de 2 meses para o fim da campanha política. Marina mostra em sua nuvem de palavras uma certa inconsistência, já que não foram ressaltadas nenhuma palavra que expressasse o sentimento geral. Porém ao analisar a classificação de suas mensagens como um crescente positivo, onde as mensagens de apoio e identificação com a candidata aumentaram e eram maioria desde o primeiro período até o último, mesmo inconsistente Marina Silva apresentou um alto índice de aceitação.

Finalizando temos o candidato Aécio Neves, que apresentou um alto índice de aceitação e adesão de sua campanha desde o primeiro período o que contrasta com a candidata Dilma Roussef. Aécio na sua nuvem de palavras ressalta palavras de apoio e identificação por parte dos usuários. Suas mensagens apontam para um forte uso de marketing e disseminação de informação sobre o mesmo. Outro ponto que chamou atenção foi a baixa taxa de rejeição,

quase inexistente na análise dos resultados do mesmo, algo variando entre 2% e 4% das mensagens capturadas.

6.1 Trabalhos futuros

Como sugestão para trabalhos futuros relacionados na mesma área, de análise de sentimento sobre candidatos, pode ser extraído dados de outras fontes, por exemplo, novas redes sociais podem ser utilizadas no escopo (Facebook, Google+, Instagram, dentre outras), isso aumentaria consideravelmente a massa de dados e daria mais confiabilidade no resultado. Outro classificador poderia ser utilizado, não somente o Naive Bayes. Outra contribuição poderia ser o uso dos modelos gerados aqui para uma classificação em tempo real das mensagens postadas na rede social Twitter sobre os candidatos numa próxima eleição, se estes ainda forem candidatos.

Com relação as mensagens coletadas, fica a sugestão de analisar percentualmente quais os valores de “rt” e “http” utilizado nas mensagens, bem como uma análise mais profunda do perfil dos usuários da rede social da qual foram coletados os dados. O que pode explicar algumas discrepâncias no resultado, visto que redes sociais tem perfis de usuários diferentes.

REFERÊNCIAS

ARANHA, C.N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional.** 2013. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2001.

BARBOSA, Glívia A. R. et al. **Caracterização do uso de hashtags do Twitter para mensurar o sentimento da população online: Um estudo de caso nas Eleições Presidenciais dos EUA em 2012.** Simpósio Brasileiro de Banco de Dados: SBBD 2013, Minas Gerais, v. 1, n. 1, p.0-0, set. 2013.

BARBOSA, Glívia A. R. et al. **Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment.** In: Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts. Austin, 2012.

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

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases.** AAAI 97, Providence, v. 1, n. 1, p.0-0, jan. 1997.

HOTHO, Andreas **A brief survey of text mining**. University of Kassel. 2005.

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

NGOC, Frederic Dang. **Using the Mahout Naive Bayes Classifier to automatically classify Twitter messages.** 2013. Disponível em: <https://chimpler.wordpress.com/2013/03/13/using-the-mahout-naive-bayes-classifier-to-automatically-classify-twitter-messages/>. Acesso em: 26 nov. 2014.

OGURI, Pedro. **Aprendizado de Máquina para o Problema de Sentiment Classification.** 2006. 52 f. Dissertação (Mestrado) - Curso de Mestrado em Informática, Departamento de Informática do Centro Técnico Científico da PUC–Rio, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006.

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

SOUSA, Giulia Luan Santos de. **TWEETMINING: Análise de opinião contida em textos extraídos do Twitter.** 2012. 66 f. TCC (Graduação) - Curso de Sistemas de Informação, Universidade Federal de Lavras, Lavras, 2012.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining: Mineração de Dados.** Rio de Janeiro: Ciência Moderna, 2009. 900 p. Tradução de Introduction to Datamining.