

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Rafael Campos Attux

**Predição dos resultados das eleições 2014 para
presidente do Brasil usando dados do *Twitter***

Uberlândia, Brasil

2017

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Rafael Campos Attux

**Predição dos resultados das eleições 2014 para
presidente do Brasil usando dados do *Twitter***

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, Minas Gerais, como
requisito exigido parcial à obtenção do grau
de Bacharel em Sistemas de Informação.

Orientador: Prof^ª. Elaine Ribeiro de Faria

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2017

Rafael Campos Attux

Predição dos resultados das eleições 2014 para presidente do Brasil usando dados do *Twitter*

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 21 de Julho de 2017:

Prof^a. Elaine Ribeiro de Faria
Orientador

Maria Adriana Vidigal de Lima
Professora

Christiane Regina Soares Brasil
Professora

Uberlândia, Brasil
2017

Dedico este trabalho a todos os professores que agregaram conhecimento a minha formação, minha orientadora que muito me ensinou, minha família, minha amada esposa e principalmente a Deus por me ajudar nos momentos mais difíceis.

Agradecimentos

A esta universidade, corpo docente, direção e administração que abriram as portas para o meu futuro, vislumbrando um horizonte superior, repleto de confiança e ética aqui presentes.

Aos meus familiares e minha esposa que pacientemente sempre me apoiou em todas as minhas decisões e é meu braço direito na caminhada da vida.

Agradeço especialmente também a minha orientadora Elaine Ribeiro de Faria, pela correção e incentivo em todos os momentos aos quais foram necessários, abrindo mão de seu precioso tempo.

Resumo

Com o crescimento da utilização das redes sociais no Brasil, abriu-se uma porta para que milhões de pessoas expressem suas emoções e sentimentos para seus amigos e seguidores, principalmente no *Twitter*. Este volume de informações podem ser aproveitados para a realização de predições sobre diversos assuntos. Este trabalho visa combinar da melhor forma possível métodos e algoritmos que possam realizar a predição de resultados das eleições brasileiras para presidente do Brasil no ano de 2014, com base em comentários realizados por usuários da rede social *Twitter* em todo o país. Neste trabalho serão coletadas amostras dos dados relativas aos principais candidatos presentes nas pesquisas eleitorais, bem como as possíveis combinações com seus partidos e vice-candidatos. Posteriormente estes dados serão tratados e classificados conforme o sentimento que ele transmite, podendo ser positivo, negativo ou até mesmo neutro. Espera-se provar a eficácia de tais métodos com resultados plausíveis e abrir novas portas para esta área de Mineração de Sentimentos em Redes Sociais, provando o grande poder que estas Redes Sociais podem ter para comprovar o sentimento de um determinado público.

Palavras-chave: Análise de Sentimentos, Mineração de Dados, Predições de Eleições, Redes Sociais, Twitter.

Lista de ilustrações

Figura 1 – Fluxograma de interação entre aplicações.	22
Figura 2 – Exemplo de Requisição manual feita na API por meio do uso do navegador.	23
Figura 3 – Fluxo OAuth controle de tokens.	24
Figura 4 – Interface Básica Requisições QVSource	25
Figura 5 – Quantidade de registros coletados por data.	34
Figura 6 – Tweets antes e após processamento de remoção de links.	35
Figura 7 – Resultados de classificações Positiva.	37
Figura 8 – Resultados de classificações Negativas.	37
Figura 9 – Classificação por Sentimentos de cada Candidato.	38
Figura 10 – Percentual de Tweets que foram Classificados.	39
Figura 11 – Resultados de classificações Neutras.	40
Figura 12 – Intenção de votos por Candidato.	40
Figura 13 – Resultado da Eleição Presidencial no Brasil de 2014.	41
Figura 14 – Acesso a Internet conforme Censo 2010.	42

Lista de tabelas

Tabela 1 – Principais Redes Sociais do Mundo e suas Características	15
Tabela 2 – Distribuição da Coleta de Dados por Candidato.	22
Tabela 3 – Exemplos de Expressões Coloquiais e suas correspondências	28
Tabela 4 – Comparativo de velocidade de leitura no QlikView.	33
Tabela 5 – Datas dos Debates das principais emissoras de TV aberta.	35
Tabela 6 – Comparativo de Intenção de Votos.	41
Tabela 7 – Distribuição de localidade por região.	43

Lista de abreviaturas e siglas

Twitter	Rede social em formato de microblogging.
Tweets	Comentários realizados por um usuário do Twitter.
Facebook	Rede social que reúne pessoas a seus amigos.
API	Interface de Programação para Aplicações.
Emoticons	Sequência de Caracteres Tipográficos.
Stop Words	Palavras que aparecem com frequência em frases e que não agregam valor para o sentimento da frase.
BI	Business Intelligence ou Inteligência de Negócios.
TSE	Tribunal Superior Eleitoral.
CSV	Comma-separated values.

Sumário

1	INTRODUÇÃO	11
1.1	Problema	12
1.2	Objetivo	12
1.2.1	Geral	12
1.2.2	Específico	13
1.3	Justificativa	13
1.4	Organização do Trabalho	13
2	REFERENCIAL TEÓRICO	15
2.1	Redes Sociais	15
2.1.1	Twitter	16
2.2	Mineração de Dados	16
2.3	Pré-processamento de Dados	17
2.4	Algoritmos de Classificação	18
2.4.1	Aprendizado de Máquina Supervisionado	18
2.4.2	Aprendizado não Supervisionado	18
2.5	Trabalhos Relacionados	19
2.5.1	Análise de opiniões expressas nas redes sociais	19
2.5.2	<i>What 140 Characters Reveal about Political Sentiment</i>	19
2.5.3	Predição de Resultado de eleição para Reitor de Universidade usando Tweets como fonte de pesquisa	20
2.5.4	<i>Twitter Sentiment Classification using Distant Supervision</i>	20
3	PROPOSTA	21
3.1	Introdução	21
3.2	Coleta de Dados	22
3.3	Pré-processamento de Dados	26
3.3.1	Remoção de Links	27
3.3.2	Eliminação de Caracteres Especiais	27
3.3.3	Linguagem Coloquial	28
3.3.4	Tradução	29
3.4	Classificação	30
3.5	Comparação entre o resultado da classificação dos tweets e o resultado das eleições	32
4	RESULTADOS E DISCUSSÕES	33

4.1	Base de Dados	33
4.1.1	Formato .QVD	33
4.1.2	Coleta de Dados	34
4.2	Normalização	35
4.2.1	Remoção de Links	35
4.2.2	Duplicados e ReTweet	36
4.3	Classificação	36
4.4	Resultados e Análises	37
4.4.1	Separação por Sentimentos	38
4.4.2	Percentual de Classificados	38
4.4.3	Intenção de Votos	40
5	CONSIDERAÇÕES FINAIS	44
5.1	Análise geral do trabalho	44
5.2	Trabalhos Futuros	44
	REFERÊNCIAS	46

1 Introdução

Atualmente existem dezenas de redes sociais muito utilizadas pelos internautas de todo o mundo, e que contém informações que se utilizadas de forma correta, podem ser muito valiosas. Segundo a empresa de consultoria ComScore, em sua pesquisa sobre as tendências do mercado digital brasileiro ([BRASILDIGITALFUTURE](#),), a rede social *Twitter* foi a terceira mais acessada pelos brasileiros em todo o ano de 2013, dentre o *ranking* de todas as redes sociais de todo o país.

Segundo o site ([FOLHASAOPAULO](#), 2017), o Brasil foi o país que obteve o terceiro maior crescimento em número de usuários ativos no *Twitter* na comparação dos anos de 2015 para 2016, alcançando 18% de crescimento neste período. Ainda conforme a reportagem, a receita da empresa cresceu 30% no país neste mesmo período, superando a média global, que foi de 14%.

Analisando a enorme importância que este meio de comunicação tem tomado, pode-se extrair de forma significativa do *Twitter* dados que revelem os sentimentos de determinados públicos que seriam complicados de se conseguir por meio de outros métodos, como por exemplo pesquisas em domicílios ou pesquisas por amostragem, como são feitas hoje em dia quando organizações querem conhecer como esta a aceitação da população em relação a determinadas marcas de produtos ou serviços.

Segundo ([CEVI](#), 2008), existem inúmeras possibilidades de aplicações em redes sociais que podem ser implementadas e estudadas em diferentes áreas de conhecimento como o mapeamento de redes profissionais como é o caso do *LinkedIn*, que proporciona aos seus usuários a possibilidade de mapear e aumentar a sua rede de contatos profissionais. Enquanto que na área de *e-Learning* existe a plataforma *Moodle* que é um sistema de gestão de cursos que possibilita a comunicação e interação entre usuários.

Outra possibilidade de implementação descrita por ([TEIXEIRA; AZEVEDO](#), 2011), é a utilização das redes sociais para analisar qual tem sido a aceitação dos usuários perante o valor de venda de determinados produtos em sua primeira semana de comercialização no mercado. Outro ponto abordado no trabalhado citado, é a análise da aceitação dos usuários das redes sociais *Twitter* e *Facebook* quanto ao lançamento de novos filmes nos cinemas.

Conforme ([CHEN; ZIMBRA](#), 2011) a análise de sentimentos em redes sociais tem como objetivo compreender e analisar as opiniões de um publico geral de possíveis consumidores, eleitores ou qualquer outro tipo de usuário que queria expressar o seu sentimento em público, seja ele negativo ou positivo.

1.1 Problema

Dentre as possíveis aplicações da análise de sentimentos em redes sociais, uma delas é em relação ao cenário político. Cada vez mais, eleitores utilizam deste meio para expor a sua satisfação ou indignação a determinado candidato ou partido eleitoral. O volume de dados gerados pelos usuários é intensificado principalmente perto das vésperas de debates eleitorais e também antes e durante do período de campanha política partidária.

Para se encomendar pesquisas eleitorais como são feitas por grandes partidos políticos e emissoras de TV, durante os períodos eleitorais, normalmente é preciso montar uma grande equipe de pesquisadores envolvendo uma logística de distribuição enorme, principalmente em países como Brasil, de grande extensão territorial. Acrescenta-se a isso o fato de que os custos de tais operações são bastante elevados.

As pesquisas eleitorais são importantes fontes democráticas de opiniões. A curiosidade em antecipar resultados é natural do ser humano, mas a principal finalidade de uma pesquisa está no contexto estratégico dos candidatos, que podem planejar ações, como mobilização para uma provável disputa de segundo turno, agradecimento aos eleitores, ou mesmo reconhecimento de uma derrota antes do resultado oficial.

Segundo ([KRAEMER; KAWAMOTO; GEROSA, 2012](#)), onde é possível ter uma ideia do quão custoso é uma pesquisa eleitoral, e como uma forma alternativa como a mineração de dados pelo *Twitter* pode ser uma fonte econômica e igualmente eficaz de conseguir predições concretas sobre o assunto. Ainda conforme o autor, são aplicados métodos de análise de sentimentos para a predição dos resultados de uma eleição para Reitor para a Universidade Tecnológica Federal do paran  (UTFPR), mostrando a viabilidade de tais processos.

  poss vel encontrar outro exemplo pr tico implementado por ([A.SHAMMA; KENNEDY; CHURCHILL, 2012](#)) onde foram analisados os dados do *Twitter* durante os debates presidenciais para o governo norte Americano no ano de 2008. O objetivo era analisar a influ ncia dos debates nas redes sociais para determinar o desempenho dos candidatos em tempo real.

1.2 Objetivo

1.2.1 Geral

Este trabalho tem como seu principal objetivo propor e avaliar uma combina  o das melhores t cnicas e m todos existentes referentes   minera  o de sentimentos em redes sociais para realizar a predi  o dos resultados eleitorais para presid ncia do Brasil no ano de 2014 com base nas informa  es extra  das do *Twitter*.

1.2.2 Específico

1. Usar uma ferramenta adequada para coleta de dados do *Twitter* referente às eleições para presidente do Brasil no ano de 2014, montando uma base de dados com um número considerável de documentos (*tweets*).
2. Pré processar a base de dados, com o objetivo de remover *tweets* indesejados, tokenizar as palavras, corrigir ou remover palavras indesejadas, limpeza de *stop words*.
3. Classificar o sentimento presente nos *tweets* como positivo, negativo ou neutro usando técnicas de classificação presentes na literatura.
4. Verificar a relação entre o sentimento dos usuários do *Twitter* em relação a cada candidato, comparando com o resultado das eleições.

1.3 Justificativa

Atualmente milhares de pessoas acessam as redes sociais para expressar seus sentimentos, principalmente no *Twitter* que é uma rede social voltada para pequenas postagens de até 140 caracteres, no qual uma pessoa pode expor para as outras que a seguem o que acharam de determinados produtos, marca, restaurante, comidas, candidatos e partidos políticos dentre outros diversos assuntos.

Essa limitação de caracteres é um dos grandes motivos para a utilização do *Twitter* como a fonte de dados para este trabalho, pois desta forma é possível formar uma base de dados mais compacta, e que transmite um forte sentimento em cada publicação de usuário, pois se houvesse um texto muito grande, a probabilidade de serem geradas ambiguidades seria maior.

Com foco neste poder de expressão, é possível explorar uma série de oportunidades, onde podemos identificar e prever comportamentos de grupos de pessoas (BOTELHO; UGIONI, 2012). Com este incrível potencial em poder descobrir inúmeras tendências, é que se justifica este trabalho, visando utilizar milhares de informações que estão disponíveis para consulta, como uma forma de rastrear as tendências do ambiente político atual.

1.4 Organização do Trabalho

É possível estabelecer uma relação entre os sentimento dos usuários do *Twitter* referente aos candidatos à presidência do Brasil e os resultados das eleições do Brasil no ano de 2014. Esse correlacionamento entre os resultados pode existir e ser equivalente.

Serão utilizados os seguintes métodos para o estudo e desenvolvimento do trabalho:

1. Estudo e escolha da melhor *API* do *Twitter*, para a extração da base de dados de *tweets* que será utilizada no trabalho.
2. Utilização das técnicas de pré-processamento da base de dados para limpeza de textos indesejáveis para o algoritmo classificador.
3. Classificação de sentimentos existentes para categorizar cada *Tweet* em positivo, negativo ou neutro, conforme informações relevantes em cada comentário.
4. Aplicação de técnicas de mineração de dados para categorizar os sentimentos dos usuários como positivo, negativo ou neutro em relação aos candidatos à presidência.
5. Comparar os resultados das eleições com os resultados obtidos no projeto.
6. Analisar a eficiência do *Twitter* para mineração de dados das eleições para presidente do Brasil no ano de 2014.

Com este trabalho concluído, serão geradas comparações positivas quanto à confiabilidade da análise de sentimentos no *Twitter* para a determinação de resultados e tendências do ambiente político brasileiro atual. Comprovar a qualidade do método utilizado ao decorrer deste projeto, levando sempre em conta o tema proposto e a sua base de dados, pois para cada outro tema diferente do deste projeto, podem haver outras soluções e técnicas melhores e mais eficientes.

2 Referencial Teórico

Para melhor entendimento do assunto proposto neste trabalho, serão discutidos alguns fundamentos das redes sociais, que são a fonte de onde serão retirados todas as informações que compõem a base de dados do projeto. Também serão apresentados os principais termos relativos a mineração de sentimentos.

2.1 Redes Sociais

Conforme a definição de (BENEVENUTO; ALMEIDA; SILVA, 2011) uma rede social online é um serviço Web que permite a um indivíduo construir perfis públicos ou semipúblicos dentro de um sistema, articular uma lista de outros usuários com os quais ele(a) compartilha conexões e visualizar e percorrer suas listas de conexões assim como outras listas criadas por outros usuários do sistema.

A Tabela 1 referencia as principais redes sociais do mundo e à quais ramos de comunicação elas se enquadram:

Tabela 1 – Principais Redes Sociais do Mundo e suas Características

Nome	Propósito	URL
Twitter	Amizades	http://twitter.com
Facebook	Amizades	http://www.facebook.com
Orkut	Amizades	http://www.orkut.com
LinkedIn	Relacionamentos Profissionais	http://www.linkedin.com
YouTube	Compartilhamento de Vídeos	http://www.youtube.com

O principal papel de uma rede social é ser um meio de comunicação utilizado por internautas para expressar os seus gostos, sentimentos, vontades, desejos, opiniões e até mesmo para influenciar outras pessoas, como é o caso de perfis de pessoas famosas e celebridades (SOUZA, 2013).

Atualmente as redes sociais tem despertado grandes interesses de empresas para a geração de propaganda, tendo em vista o crescente número de usuários e que podem ser facilmente influenciados por anúncios e propagandas pagas. Existe também o mercado de jogos que viram febre nesses meios de comunicação, tornando-as ainda mais interessantes do ponto de vista financeiro para pequenas e grandes empresas ganhar dinheiro (SOUZA, 2013).

Ainda segundo (BENEVENUTO; ALMEIDA; SILVA, 2011) serão apresentadas a seguir algumas das principais funcionalidades oferecidas pelas redes sociais e que são de

suma importância para conhecimento deste trabalho, e que talvez sejam a estratégia de tornar estas plataformas em um meio de comunicação tão eficaz:

- Perfis de usuários: Uma página individual, que oferece a descrição de um membro.
- Atualizações: Formas efetivas de ajudar usuários a descobrir conteúdo.
- Comentários: Usuários comentam o conteúdo compartilhado por outros.
- Avaliações: O conteúdo compartilhado por um usuário pode ser avaliado por outros usuários.
- Listas de Favoritos: Permite que usuários selecionem e organizem seu conteúdo.

2.1.1 Twitter

Dentro do contexto das redes sociais se destaca o *Twitter*, que é um serviço de micro-blogging criado para descobrir o que está acontecendo em qualquer hora ou lugar do mundo (BIFET; FRANK, 2010). Um dos grandes diferenciais desta plataforma e que talvez seja o grande motivo do seu sucesso, é a limitação da quantidade de caracteres que podem ser postados a cada mensagem, cuja é de 140. Tais mensagens criadas pelos usuários recebem o nome de *tweets*.

Grandes empresas especializadas em propagação de notícias como a Rede Globo e demais emissoras de TV e jornais possuem perfis no *Twitter* para auxiliar a transmissão de notícias, pois esta ferramenta atualmente tem se mostrado a mais eficaz quando se diz respeito à velocidade com que a informação é enviada e propagada.

Essa limitação de caracteres é um dos grandes motivos para utilizarmos o *Twitter* como a nossa fonte de dados para este trabalho, pois desta forma temos uma base de dados mais compacta, e que transmite um forte sentimento em cada publicação de usuário, pois se houvesse um texto muito grande, a probabilidade de serem geradas ambiguidades seria maior.

2.2 Mineração de Dados

A mineração de dados (do inglês, *Data Mining*), tem como principal utilidade a varredura de grande quantidade de dados a procura de padrões e detecção de relacionamentos entre informações gerando novos subgrupos de dados. A formação de subgrupos de dados é feita pela mineração de dados por meio da execução de algoritmos capazes de conhecer e aprender mediante a varredura dessas informações (BOTELHO; UGIONI, 2012).

A partir destes subgrupos que são gerados pela mineração de dados, é que podemos chegar a conclusões sobre o que cada informação contida nesses agrupamentos podem revelar sobre o contexto analisado em questão.

Com base em análises públicas em ([MICROSOFTARTICLES, 2017](#)) existem 6 principais etapas resumidas abaixo a serem seguidas para conseguir a realização de uma mineração de dados com sucesso:

1. Definição do Problema: Definir claramente o problema e considerar maneiras de os dados serem utilizados para fornecer respostas para ele.
2. Preparação dos Dados: Consolidar e limpar os dados, pois muitas vezes eles podem estar espalhados, ou em formatos diferentes, realização de um Pré-Processamento da base.
3. Exploração dos Dados: Compreender os dados para tomar decisões apropriadas ao criar os modelos de mineração. Conhecer a fundo com o que tipo de informações estamos trabalhando, e o que podemos encontrar de dificuldades.
4. Criação de Modelos: Definir as colunas de dados que serão usadas ao criar uma estrutura de mineração.
5. Exploração e Validação: Unir modelos criados durante a fase de teste e validação final para teste de eficiência.
6. Implantação e Atualização: Implantar os modelos que tiveram o melhor desempenho em um ambiente de produção.

2.3 Pré-processamento de Dados

A etapa de pré-processamento, no processo de descoberta de conhecimento, compreende a aplicação de várias técnicas para captação, organização, tratamento e a preparação dos dados. Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados ([ALGORITMOS, 2014](#)).

Este processo é extremamente importante para o projeto, pois é ele o responsável pela formatação e estruturação dos dados que compõem toda a base de dados, garantindo que ao realizar a etapa de Classificação, todos os dados possam ser interpretados e categorizados corretamente pelo algoritmo classificador. Caso haja alguma anomalia prejudicial na base, ela terá de ser corrigida nesta etapa.

Para este projeto, são necessárias aplicações de várias técnicas para o pré-processamento dos dados, como remoção de *links*, eliminação de caracteres especiais, adequação da linguagem coloquial para a linguagem formal e também a remoção de registros duplicados.

2.4 Algoritmos de Classificação

Atualmente existem vários tipos de algoritmos de classificação de sentimentos utilizados por desenvolvedores e pesquisadores, alguns dos mais importantes estão listados nesta seção.

2.4.1 Aprendizado de Máquina Supervisionado

Segundo (LIU; ZHANG, 2012) uma das formas mais eficientes para a utilização deste tipo de algoritmo é por meio do método conhecido como de TF-IDF, que tem o objetivo de classificar o sentimento de uma frase ou um documento com base na contagem da frequência de cada termo, ou seja, são analisadas quantas vezes foram repetidas cada palavra e qual a importância de cada uma para o tema em questão a ser classificado. Quanto mais vezes uma palavra é utilizada em um documento, maior será o seu peso, aumentando de forma proporcional à sua frequência, sempre levando em consideração o volume total de palavras do documento.

Um dos algoritmos de classificação mais conhecidos e utilizados é o de *Naive Bayes*, que tem como função realizar o cálculo probabilístico da classificação de termos dos sub-grupos com base no valor atribuído sobre uma determinada classe, utilizando um modelo de treinamento previamente importado.

Ainda conforme (LIU; ZHANG, 2012), este processo tem como objetivo a construção de uma classificação prévia de sentimentos para palavras e frases mais utilizadas no contexto desejado, ou seja, é feita uma etapa de treinamento do algoritmo classificador para que o mesmo seja capaz de analisar os sentimentos desejados com base nas informações que o mesmo já possui a respeito de determinado assunto ou tema.

2.4.2 Aprendizado não Supervisionado

Neste processo, são extraídos das frases os adjetivos ou advérbios, pois os mesmos são bons indicadores de sentimentos. Neste momento são definidos quais serão os n-Gramas utilizados no processo.

Segundo (VILELA, 2011), n-Grama é uma sequência de n itens dentro de uma frase. Os itens podem ser palavras, letras, sílabas, classificação gramatical das palavras, ou qualquer outra base. Um n-grama de tamanho 1 é chamado de unigrama, de tamanho 2, de bigrama, de tamanho 3 é chamado de trigrama, de 4 em diante é n-grama.

Ao analisar a frase "Os alunos são inteligentes", é possível criar uma cadeia de bigramas composta pelos termos: "# Os", "os alunos", "alunos são", "são inteligentes", "inteligentes #".

Com base no treinamento previamente fornecido pelo usuário ao algoritmo, o mesmo é capaz de calcular a importância de cada termo pesquisado com base na sua posição de ocorrência na frase.

2.5 Trabalhos Relacionados

Com base no grande potencial de informações em que se pode obter de uma rede social, vários estudiosos estão sempre buscando extrair o máximo de informação possível deste meio de comunicação. É um universo que está sempre em constante renovação e expansão. Este capítulo reúne alguns trabalhos relacionados com o tema de análise de sentimentos e também alguns casos em que o tema aplicado é o mesmo deste trabalho, a política.

2.5.1 Análise de opiniões expressas nas redes sociais

Este trabalho de (TEIXEIRA; AZEVEDO, 2011) tem como objetivo analisar como determinados produtos são percebidos pelos usuários das redes sociais *Twitter* e *Facebook* e verifica a relação com os valores de comercialização obtidos na sua primeira semana no mercado.

Na construção da base do projeto, foram analisados dez filmes e as respectivas mensagens recolhidas nos 7 dias anteriores às suas estreias. Foram coletadas mais de 400 mil publicações. Outro dado que foi levado em conta é o valor de Bilheteria.

Um dos problemas encontrados neste trabalho foi o da realização do processo de análise dos termos na língua portuguesa, tendo em vista que todos os comentários analisados foram somente com base em publicações em inglês e sobre filmes norte americanos.

2.5.2 *What 140 Characters Reveal about Political Sentiment*

Com foco nas eleições federais do parlamento da Alemanha no ano de 2009, segundo (TUMASJAN; SPRENGER; WELPE, 2010) foram coletados pouco mais de 104 mil *tweets* durante o período de vésperas eleitorais. Um dos pontos deste trabalho é a correlação do volume de publicações e compartilhamentos com os resultados das eleições. Ao analisar todos os partidos, houve uma margem de erro de apenas 1.65 %, ou seja apenas com base nesta informação, já é possível extrair vários indicadores.

A conclusão de (TUMASJAN; SPRENGER; WELPE, 2010) é que o *Twitter* é realmente usado como um meio para discussão política. Ao analisar apenas o volume de

tweets, é possível perceber que eles refletem as preferências dos eleitores e se aproximam das pesquisas eleitorais tradicionais.

2.5.3 Predição de Resultado de eleição para Reitor de Universidade usando Tweets como fonte de pesquisa

O trabalho de (KRAEMER; KAWAMOTO; GEROSA, 2012) tem como objetivo a predição de resultados para o cargo de Reitor da Universidade Tecnológica Federal do Paraná(UTFPR), tendo em vista que em tais situações não ocorrem pesquisas eleitorais prévias, o autor destaca a necessidade da tentativa de antecipar possíveis resultados. Neste caso em específico, existe uma divisão dos usuários em servidores e estudantes, onde cada um recebe um determinado peso de importância.

Como o volume de dados é referente a um conjunto pequeno, foram coletados 744 mensagens utilizando a *API* do *Twitter* e o processo de classificação das mensagens foi feito de forma manual, avaliando o sentido semântico para indicar a polarização de cada *tweet*.

2.5.4 *Twitter Sentiment Classification using Distant Supervision*

Este relatório técnico aborda a base da estrutura para o funcionamento da aplicação *Sentiment140* disponível em <www.sentiment140.com> que é utilizada como um algoritmo classificador de sentimentos especializado em mensagens do *Twitter*. Conforme (GO; BHAYANI; HUANG, 2013), a base desta aplicação foi treinada previamente com a utilização de termos positivos e negativos do *Twitter* para melhor desempenho do algoritmo, o qual é baseado em aprendizagem de máquina.

O autor se baseia na premissa de que a realização do treinamento do algoritmo de aprendizagem de máquina é mais eficiente quando são considerados *tweets* que possuem *emoticons* em seu contexto. Para tal treinamento, foi utilizada uma base contendo 800 mil *tweets* com *emoticons* negativo e 800 mil *tweets* com *emoticons* positivos, totalizando 1.600.000 *tweets*. Nesta base não foi definido nenhum tema específico como abordado em outros trabalhos, foram levados em conta vários temas como, produtos, filmes, personalidades, eventos e etc.

Ainda conforme (GO; BHAYANI; HUANG, 2013) foram utilizados os três principais algoritmos de aprendizagem de máquina atuais (*Naive Bayes*, *Maximum Entropy* e *Support Vector Machines*) e suas utilizações com unigramas, bigramas e ambos. A conclusão é que no geral todas as classificações foram satisfatórias para a base de treinamento, porém a combinação entre unigramas unidos com bigramas para o *Maximum Entropy* foi a que obteve melhores resultados, com precisão de 83%.

3 Proposta

3.1 Introdução

Este capítulo apresenta em detalhes a proposta deste trabalho, que consiste no estudo do uso de redes sociais para predição do resultado das eleições brasileiras do ano de 2014 para o cargo de presidente do país. Para tal estudo, a rede social utilizada foi o *Twitter*, que apresenta importantes características, tais como permitir que os seus usuários expressem o que estão pensando em pequenos textos para que todos os demais usuários vejam, compartilhem e opinem sobre determinado assunto, além de ser uma ferramenta de fácil acesso e utilização gratuita.

Todas as informações necessárias para o desenvolvimento do trabalho e a geração de dados para análise precisam ser coletadas de forma estruturada e concreta para que não haja nenhum tipo de confusão ou indução ao erro, conforme encontra-se descrito na Seção 3.2.

Os registros do *Twitter* antes de serem analisados precisam ser submetidos a um processo de pré-processamento, que é uma etapa onde serão retiradas as informações irrelevantes para a classificação, como por exemplo caracteres especiais indesejados e também erros de escrita informal. Esta etapa é de extrema importância para o resultado final, pois é nela que podem ser encontrados e retirados caracteres especiais ou erros de ortografia que são prejudiciais para o algoritmo classificador. Os procedimentos utilizados nesta etapa, estão descritos na Seção 3.3.

Após ter feito todos estes passos, todos os dados devem ser enviados para uma próxima etapa, que é a Classificação de tudo aquilo que foi coletado e pré-processado. Durante este processo é importante lembrar que cada candidato terá suas informações processadas de forma independente um do outro, ou seja, ao processar o sentimento relativo ao candidato X, não é levado em conta o sentimento do mesmo usuário em relação ao candidato Y. Todos os detalhes da ferramenta e da maneira que ela foi utilizada para Classificação podem ser encontrados na Seção 3.4

A última etapa deste processo, ocorre na Análise dos Resultados obtidos após a Classificação da base coletada. Neste momento são levados em conta também os dados reais das eleições a fim de se verificar se existe uma relação entre o resultado das eleições e o sentimento expresso na amostra de *tweets* coletados nesse período. Todos os resultados foram analisados conforme descrito na Seção 3.5.

3.2 Coleta de Dados

A coleta de dados da rede social *Twitter* foi realizada diariamente durante todo o período de campanha eleitoral para as eleições em primeiro turno, entre os dias 27 de Agosto e 03 de Outubro do ano de 2014. As informações foram capturadas por meio de solicitações manuais feitas pela intermediação da *API* descrita no tópico 1. Tais consultas foram feitas de maneira incremental para cada candidato, visando a estruturação de uma base sólida e sem registros duplicados, para que todo o processo posterior tivesse um grande nível de confiabilidade e segurança.

As consultas à base de dados original do *Twitter* foram feitas com base nos nomes e/ou sobrenomes de cada candidato. Foram coletados um total de 330.881 registros, contendo os textos originais publicados, identificador único de cada registro, o candidato à qual cada registro se refere, localidade de cada usuário, data e hora da publicação no *Twitter* e vários outros dados que possam ser relevantes no futuro.

Tabela 2 – Distribuição da Coleta de Dados por Candidato.

Candidato	Termos Pesquisados	Quantidade de Registros
Aécio Neves	aecio aecio neves	116.637
Dilma Rousseff	dilma dilma rousseff	105.972
Marina Silva	marina silva	77.003
Pastor Everaldo	pr. everaldo pastor everaldo	31.269

A tabela 2 mostra a distribuição da quantidade de registros coletados por candidato e os termos utilizados para a pesquisa.

Para a execução da tarefa de coleta de dados, as seguintes ferramentas foram utilizadas conforme detalhado no fluxograma da Figura 1 e respectivamente explanadas ao decorrer desta seção.

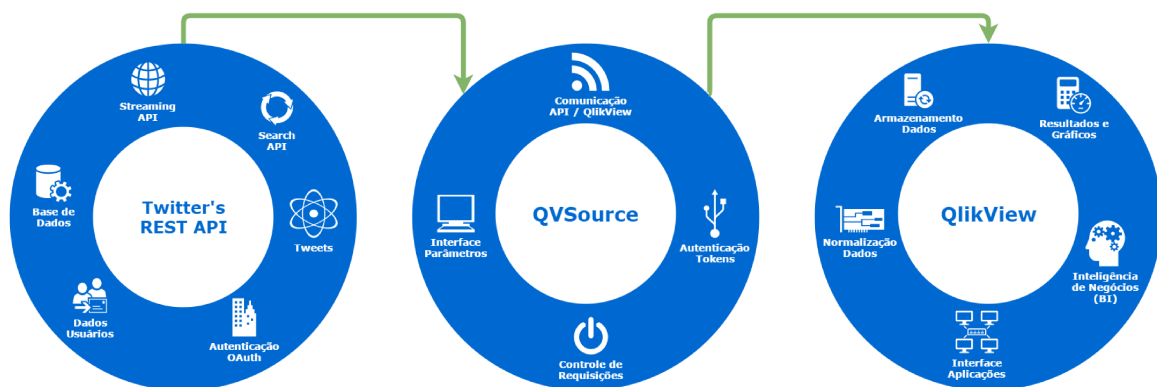


Figura 1 – Fluxograma de interação entre aplicações.

1. Twitter's v1.1 REST API

Esta ferramenta visa fornecer acesso programático a ler e escrever os dados da rede social *Twitter* (RESTAPIS, 2014). Sua função é de estabelecer uma conexão entre o usuário da *API* e o banco de dados do *Twitter*, retornando assim todas as informações e dados sobre os *tweets* ou qualquer outra informação disponível para consulta. Esta ferramenta é de uso gratuito, porém ela possui algumas restrições conforme esta descrito a seguir neste capítulo. Toda a interface de comunicação e requisição pode ser feita de várias maneiras, tais como, manual, usando o navegador por meio de *URL* (ver figura 2), ou por meio de qualquer outro programa que seja capaz de interpretar os resultados gerados pela *API*, como é o caso deste trabalho, no qual foi utilizada a ferramenta *QVSource*, descrita no item 2.



Figura 2 – Exemplo de Requisição manual feita na API por meio do uso do navegador.

A *API* do *Twitter* pode ser dividida em três partes (TIPOSAPIS, 2016). Este projeto utiliza somente as duas primeiras *APIs* da lista a seguir:

- a) REST API: Esta é a principal *API*. É a que possui mais serviços e destina-se à manipulação dos dados dos usuários e conexões entre eles, bem como ao envio de mensagens. É importante observar que todas as *APIs* utilizam REST, mesmo que apenas esta tenha REST no nome.
- b) Search API: A *API* de busca disponibiliza serviços para pesquisa de mensagens e usuários.
- c) Streaming API: Esta *API* possui uma arquitetura um pouco diferenciada das duas primeiras. Destina-se a gerar conexões persistentes para troca de informações de forma síncrona. Isso é útil em sistemas *desktop*, ou que necessitem atualizar o histórico de mensagens constantemente.

Outra funcionalidade da *API* que deve ser considerada é o seu método de autenticação para a liberação de acessos. Esta autorização é feita por meio do protocolo OAuth que é um método utilizado para garantir o acesso às informações desejadas dos *tweets* de forma segura e autenticada por aplicações externas.

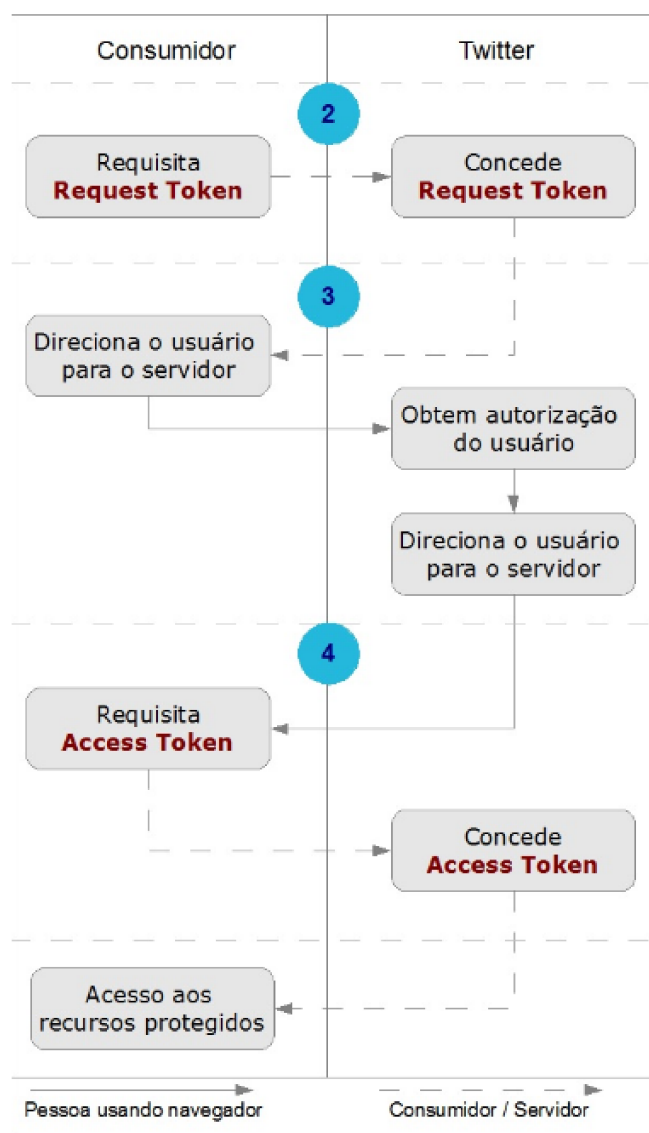


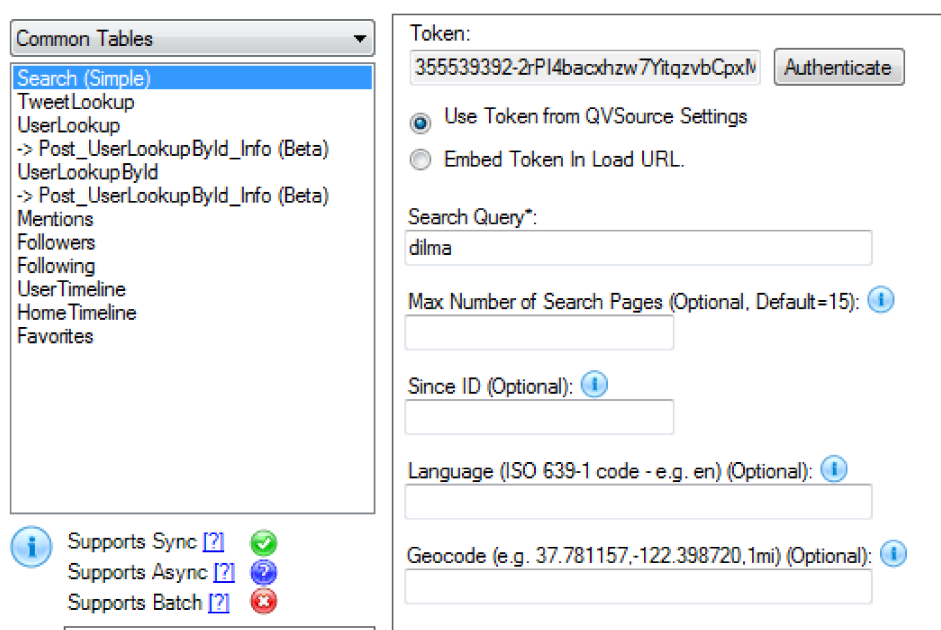
Figura 3 – Fluxo OAuth controle de tokens.

Antes de iniciar o fluxo de autenticação OAuth, é necessário realizar o primeiro passo que é o registro da aplicação na base de dados do *Twitter*. A ilustração da figura 3 do fluxo de autenticação OAuth, inicia no passo dois, onde são gerados os Tokens de acesso e liberação aos dados desejados pelo usuário para cada transação. Uma das funcionalidades da ferramenta listada a seguir, o *QVSource*, é facilitar a interface de autenticação, pois a mesma já possui todos os métodos implementados de forma visual, simples e prática.

2. QVSource v1.5.4.6

Este extrator é uma ferramenta independente utilizada como um complemento de interface ao REST API, pois é ele quem auxilia no paralelismo das requisições enviadas à API e também gerencia as autenticações de acesso. O seu resultado é uma saída otimizada de informações, já estruturada para a aplicação QlikView,

listada na sequencia. Ao utilizar esta ferramenta é necessário sincronizar uma chave de acesso única chamada de *Token*, que é uma credencial gerada pela *REST API* e que deve ser inserida na interface de comunicação *QVSource* para que possa ser liberada para esta aplicação o consumo de requisições e respostas para conseguir os *tweets* que irão compor a base de dados utilizada neste projeto. Este *Token* é obrigatório, pois ele foi criado a fim de controlar e evitar que pessoas ou processos consumam um grande tráfego de informações, tornando o acesso a esta rede social mais lento e até mesmo ocasionando uma possível queda do site.



The image shows a web interface for QVSource. On the left, there is a 'Common Tables' dropdown menu with a list of tables: Search (Simple), TweetLookup, UserLookup, -> Post_UserLookupById_Info (Beta), UserLookupById, -> Post_UserLookupById_Info (Beta), Mentions, Followers, Following, UserTimeline, HomeTimeline, and Favorites. Below this list are three status indicators: 'Supports Sync' with a green checkmark, 'Supports Async' with a blue question mark, and 'Supports Batch' with a red X. On the right, there is a 'Token' field containing a long alphanumeric string, followed by an 'Authenticate' button. Below the token field are two radio buttons: 'Use Token from QVSource Settings' (selected) and 'Embed Token In Load URL'. Further down are several input fields: 'Search Query*' with the text 'dilma', 'Max Number of Search Pages (Optional, Default=15):', 'Since ID (Optional):', 'Language (ISO 639-1 code - e.g. en) (Optional):', and 'Geocode (e.g. 37.781157,-122.398720,1mi) (Optional):'. Each of these fields has a blue information icon to its right.

Figura 4 – Interface Básica Requisições QVSource

A criação do *Token* é bem simples (ver Figura 4), bastando apenas ser validada dentro do próprio *QVSource* junto à uma conta válida do *Twitter*. Uma vez criado o *Token*, o extrator em questão fica responsável pelo seu gerenciamento.

A utilização desta ferramenta pode ser feita tanto de forma isolada, por meio de sua interface simplificada no qual os resultados das consultas são exibidos em uma tabela dentro do próprio extrator, ou de forma integrada, onde será gerado um script com códigos para ser utilizado na aplicação *QlikView*, a qual foi utilizada para este projeto.

3. QlikView v11.20 SR5

O *Qlikview* pode ser baixado de forma gratuita em www.qlik.com/us/products/qlikview para uma versão de avaliação onde o usuário só poderá abrir aplicações que ele mesmo criou em sua própria máquina, ou também de forma paga, na qual a

única diferença é que o desenvolvedor poderá abrir qualquer tipo de aplicação criada por outras pessoas. Este trabalho utiliza a versão paga que possui uma licença própria, porém para este propósito, qualquer uma das duas opções teriam o mesmo resultado desejado.

Esta é uma aplicação de BI que armazena e processa todas as informações recolhidas por todos os itens listado acima, bem como é a responsável por todas as tratativas e análises realizadas neste projeto. Sua aplicação se dará em praticamente todas as etapas deste projeto, desde a coleta, pré-processamento, tratativas e estruturação, importação e exportação de informações da base e também na análise de resultados.

O *Qlikview* foi escolhido para este trabalho devido a sua fácil comunicação com o extrator *QVSource*, pela sua performance e capacidade de armazenar e recuperar os dados gerados, além de ser uma ótima ferramenta de geração de relatórios e indicadores para análise de resultados dentro do tema escolhido.

Devido a uma restrição da *API* utilizada, na qual é limitada à 1.800 registros à cada 15 minutos ([RESTAPIS, 2014](#)), a coleta teve de ser feita em diversos períodos do dia, entre as datas 27 de agosto e 03 de outubro de 2014. Caso não houvesse tal restrição em limitar a quantidade de informações em um determinado intervalo de tempo, seria possível executar o processo de captura apenas uma vez por dia capturando tudo aquilo que fosse novo para a base de dados.

Para conseguir realizar a quantidade necessárias de consultas, foi feita a criação de uma base incremental no *QlikView*, onde após cada requisição foi salvo o número identificador do último *tweet* recolhido. Assim, as próximas requisições eram sempre feitas pegando tudo que era mais recente do que o último registro salvo. Toda a base foi salva em um arquivo específico e interno do *QliKView*, o qual está sendo utilizado como uma espécie de banco de dados de todo o projeto.

3.3 Pré-processamento de Dados

Todas as etapas do processo para a normalização da base de dados foram implementadas na ferramenta *QlikView*. Para todos os registros existentes na base de dados, foram realizadas várias tratativas de normalização de registros, com o intuito de garantir a melhor classificação possível.

As implementações foram feitas utilizando a combinação de várias fórmulas já existentes no *QlikView*. Durante o pré-processamento são feitas várias cargas de dados a fim de replicar cada tratativa que foi feita para todos os registros da base.

3.3.1 Remoção de Links

Remoção de todos os links existentes nos comentários, tendo em vista que eles não são importantes para a classificação, e podem vir a atrapalhar o processo.

É comum no *Twitter* que um usuário utilize imagens para complementar a sua publicação. Quando este recurso é utilizado, a imagem é convertida em um endereço de URL e anexado ao *tweet*, portanto neste caso deve ser removido também todos os links existentes. Foi implementada no código do *QlikView* uma regra onde ao encontrar no *tweet* algum token que tenha a informação "http", o mesmo é substituído na frase por um espaço em branco

Exemplo :

- Texto Original :
 - Comunicadores lançam manifesto de apoio à reeleição de Dilma Rousseff - Portal Vermelho <http://t.co/ZUMBbm3vwW>
- Texto Pós-Processado :
 - Comunicadores lançam manifesto de apoio à reeleição de Dilma Rousseff - Portal Vermelho

3.3.2 Eliminação de Caracteres Especiais

Nesta etapa de normalização, foram eliminados qualquer tipo de caracteres especiais que não trariam valor ao sentimento desejado, evitando assim que a base fosse poluída por caracteres indesejados que poderiam vir a atrapalhar todo o processo em qualquer uma de suas etapas, principalmente na tradução da língua portuguesa para a inglesa.

Para tal processo ser realizado, o melhor meio identificado, foi montar um lista com os possíveis caracteres desejados pelo sistema, de forma a montar um universo único, sendo que qualquer outro caractere que não estivesse pré-determinado não entraria no universo desejado, tendo em vista que existem mais de cinco mil caracteres possíveis e seria impossível listar todos eles para o sistema.

Exemplo :

- Texto Original :
 - Eu como "evangélica"(ai detesto esse rótulo) jamais votaria nesse Pr Everaldo ou em qq quer outro que governe apenas para
- Texto Pós-Processado :

- Eu como evangélica ai detesto esse rótulo jamais votaria nesse Pr Everaldo ou em qq quer outro que governe apenas para

3.3.3 Linguagem Coloquial

Segundo o artigo do site ([TODAMATERIA, 2017](#)), a linguagem coloquial compreende a linguagem informal, ou seja, é a linguagem cotidiana que utilizamos em situações informais, por exemplo, na conversa com os amigos, familiares, vizinhos, dentre outros. Quando utilizamos a linguagem coloquial decerto que não estamos preocupados com as normas gramaticais, e por isso, falamos de maneira rápida, espontânea, descontraída, popular e regional com o intuito de interagir com as pessoas.

Esta linguagem que é comum em redes sociais, foi preparada para ser transformada em linguagem Culta, a fim de evitar erros principalmente na etapa de tradução e classificação de sentimentos.

Para este processo, é necessária a ajuda de algumas pré-classificações de possíveis e mais comuns tipos de expressões popularmente utilizadas na língua portuguesa, principalmente, nas redes sociais.

Neste processo foi utilizada uma técnica conhecida como "tokenização", que é a separação das palavras de uma frase. Ao analisar cada palavra separadamente, é possível a aplicação das correções ortográficas e de linguagem coloquial.

Para saber quais palavras deveriam ser trocadas, foi montada uma lista com o principais *tokens* que foram mais repetidos e que estavam errados.

Na Tabela 3 estão listadas apenas alguns exemplos para melhor entendimento do que foi feito, com suas respectivas origens e correspondências.

Tabela 3 – Exemplos de Expressões Coloquiais e suas correspondências

Expressão	Correspondência
Vc	Você
Hj	Hoje
Pq	Porque
Tb	Também
Aff	Bravo
kk (ou mais seguidos)	Risos
HaHa (ou mais seguidos)	Risos
HeHe (ou mais seguidos)	Risos

3.3.4 Tradução

Tradução da base de dados para a língua Inglesa, pois o algoritmo classificador que é utilizado, esta preparado para receber palavras e frases apenas nesta linguagem. Neste caso, é utilizado o tradutor de idiomas Google Translate, que conforme (GOOGLETRANSLATE, 2016) é um serviço gratuito do Google que traduz instantaneamente palavras, frases e páginas da Web entre o inglês e mais de 100 outros idiomas.

Exemplo :

- Texto Original :
 - Em entrevista ao Bom dia Brasil, a presidenta Dilma Rousseff e Miriam Leitaο discordaram sobre medidores economicos
- Texto Pós-Tradução :
 - In an interview with Good Morning Brazil, President Dilma Rousseff and Miriam Leitaο they disagreed on economic gauges

Durante o processo de tradução da base de dados da língua portuguesa para o inglês, a base em geral foi quebrada em diversos arquivos menores e enviados para a plataforma de tradução, sendo o resultado deste processo arquivos recebidos já na língua inglesa. Este processo de envio e recebimento da tradução da base de dados foi feito de forma manual, com planilhas do Excel com os devidos conteúdos a serem traduzidos. A importação das planilhas foi feita por meio da utilização da interface disponibilizada aos usuários pela ferramenta Google Translate, que pode ser acessada em sua página principal disponível em <<https://translate.google.com.br/?hl=pt-BR>>.

Estes arquivos foram posteriormente unificados em uma única base novamente para dar continuidade à próxima etapa dentro da aplicação do *QlikView*. É necessário realizar esta quebra de arquivos devido à uma limitação da própria plataforma de tradução, onde são permitidas a tradução de apenas 250 linhas por instância de solicitação.

Neste momento, também foi gerado para cada *tweet* da base um identificador único, na forma de uma chave, para que durante toda a transição de informação entre aplicações e plataformas seja possível identificar e rastrear cada informação de forma individual, analisando os efeitos positivos ou negativos de cada etapa deste processo.

Após a conclusão com sucesso desta etapa, é possível partir do principio de que a base está sólida, estruturada e confiável para ser enviada à Classificação, descrita na seção 3.4. Pois se não houver esta tratativa, a chance do algoritmo classificador errar durante a classificação seria bem mais alta.

3.4 Classificação

Na etapa de Classificação da base de dados coletada, são analisados o sentimento que cada *tweet* transmite, se ele é positivo, negativo ou neutro. Com base no resultado final gerado após esta etapa, é possível extrair as informações e os dados necessários para a próxima etapa de Análise de Resultados, onde são comparados todos os resultados obtidos.

Para separar essas três polaridades de forma satisfatória, é utilizado neste trabalho uma ferramenta que é capaz de classificar os sentimentos de *tweets*, cuja o foco é justamente voltada para a rede social *Twitter*.

A aplicação *Sentiment140* está disponível para acesso gratuito através da sua página oficial ([SENTIMENT140, 2017](#)) a qual permite descobrir qual o sentimento dos usuários a respeito de uma determinada marca, produto ou algum assunto específico no *Twitter*. Além da interface fácil e intuitiva que é utilizada na sua página principal, o *Sentiment140* disponibiliza também o seu próprio algoritmo de classificação para ser utilizado por desenvolvedores através de sua *API*, a qual é utilizada neste trabalho.

Conforme o artigo de ([GO; BHAYANI; HUANG, 2013](#)) o autor explica detalhadamente todo o processo de desenvolvimento e testes do algoritmo classificador que foi construído com base em fórmulas de aprendizado de máquina.

Os itens a seguir detalham como funciona a divisão da *API* do *Sentiment140*. Para este projeto, a utilização do terceiro item (Bulk Classification Service (CSV)) é a mais adequada.

- Sentiment 140 API

1. *Bulk Classification Service (JSON)*

Este método utiliza o envio de requisições de classificação por meio de protocolo *HTTP POST* e o retorno dos dados em lote classificados é através da linguagem *JSON*.

Neste caso, o envio dos dados é feito em lote, vários *tweets* em uma mesma solicitação.

2. *Simple Classification Service (JSON)*

O processo deste método é bem similar ao item anterior, porém a diferença é que o envio dos dados é feito individualmente, ou seja, um *tweet* de cada vez.

Os protocolos de comunicação *HTTP POST* (envio) e *JSON* (recebimento) são os mesmos.

3. Bulk Classification Service (CSV)

Este é o método escolhido, que utiliza como parâmetro de envio um arquivo de texto contendo um bloco de *tweets* a serem executados.

O retorno deste processo é um simples arquivo CSV contendo em cada linha o *tweet* informando e sua respectiva classificação em 0 (negativo), 2 (neutro) e 4 (positivo).

Para a utilização desta ferramenta não é necessária nenhuma fase de teste ou de treinamento, pois a mesma já possui uma grande base de *tweets* que foram utilizados para o treinamento do algoritmo classificador. Sendo assim, esta etapa pode ser desconsiderada pelos usuários.

A execução da *API* é feita pelo próprio terminal de comandos do sistema operacional e é feita de forma bem simples, utilizando apenas dois parâmetros base que são diretório e *query*.

O diretório é o caminho físico onde está armazenado o arquivo de texto contendo as frases que devem ser classificadas, onde podem existir no máximo 10.000 linhas, conforme restrição já explicada anteriormente.

O segundo parâmetro (*query*), é uma função utilizada para informar ao algoritmo classificador sobre quem é a pessoa da frase a qual é preciso saber o sentimento. Esse parâmetro é essencial em casos de *tweets* que contém termos de dois ou mais candidatos ao mesmo tempo, como no exemplo abaixo :

Exemplo de *Tweet*: Eu amo a Marina e odeio a Dilma.

Neste caso essa informação pertencerá a dois candidatos diferentes, e cada uma das instâncias de processamento terá resultados diferentes, sendo positivo para a Marina e negativo para a Dilma, pois o algoritmo saberá exatamente quem é a pessoa da frase a qual desejamos descobrir o sentimento do usuário.

Abaixo estão listadas as seguintes chamadas da *API* que foram executadas para esta etapa, e seus respectivos candidatos:

- `curl -data-binary @C:/dilma.txt "http://www.sentiment140.com/api/bulkClassify?query=dilma"`
- `curl -data-binary @C:/aecio.txt "http://www.sentiment140.com/api/bulkClassify?query=aecio"`
- `curl -data-binary @C:/marina.txt "http://www.sentiment140.com/api/bulkClassify?query=marina"`
- `curl -data-binary @C:/everaldo.txt "http://www.sentiment140.com/api/bulkClassify?query=everaldo"`

O parâmetro que sempre é passado fixo *-data-binary* serve para informar ao comando para preservar a nova linha de caracteres informada no conjunto de dados que foi enviado, garantindo assim a formatação e devolução correta dos dados no retorno.

Após cada uma das quatro execuções, o retorno são todos os dados informados exibidos na própria tela de comando do sistema operacional, já classificados com os números 0, 2, 4 e prontos para serem colados diretamente em uma planilha Excel para análise futura de resultados.

3.5 Comparação entre o resultado da classificação dos tweets e o resultado das eleições

Após as etapas anteriores, todos os resultados obtidos são tratados e preparados para a geração de informações pela aplicação do *QliKView*. Os dados e gráficos são gerados com base nas fórmulas desejadas para que os resultados finais do projeto possam ser analisados e comparados com os dados reais dos votos da eleição presidencial de 2014 no Brasil.

Os resultados oficiais servem como base para a análise de possíveis acertos e falhas do projeto e quais são os fatores positivos ou negativos que afetaram a predição dos resultados.

Novamente a aplicação *QliKView* tem um papel fundamental, pois com ela é possível de forma fácil e intuitiva a criação de gráficos e indicadores, armazenamento dos dados e também em um ambiente de produção real o reprocessamento de informações de forma periódica em curtos intervalos, como por exemplo, de hora em hora.

O objetivo principal é demonstrar através dos fatores identificados durante o projeto quais são as dificuldades atuais nesta área de pesquisa e construir uma base para que a predição de resultados eleitores nas redes sociais sejam uma realidade em breve.

As classificações neutras não são levadas em contas para o cálculo final, pois para esta análise específica não são necessárias.

Foram feitas análises a respeito da localidade de cada usuário para a comparação com os resultados eleitorais por cada região. Como este campo pode ser preenchido com qualquer informação pelo usuário, ou seja, é um campo aberto no cadastro do perfil, foi feito de forma manual uma análise e separação de cada município para então ser possível fazer o seu relacionamento no *QliKView* por região e conseguir os indicadores desejados.

4 Resultados e Discussões

Esse capítulo apresenta os resultados obtidos durante cada etapa deste projeto e tem como objetivo analisar os acertos e dificuldades enfrentadas e identificar pontos críticos.

4.1 Base de Dados

4.1.1 Formato .QVD

A base de dados utilizada neste projeto é toda implementada em uma estrutura própria da aplicação *QlikView*, que são os arquivos do tipo .QVD. Este formato foi escolhido por ter uma alta taxa de velocidade de leitura e escrita, a qual foi um critério decisivo para este projeto, devido ao alto volume de informações da base de dados e a constante necessidade de operações como exclusão, inserção incremental e alterações.

Conforme informações de (QLIK, 2017), o tipo .QVD é um formato *QlikView* nativo e pode ser gravado e lido apenas pelo *QlikView*. O formato de arquivo é otimizado para agilização na leitura de dados de um script do *QlikView* e, ao mesmo tempo é compacto. A leitura de dados de um arquivo QVD é geralmente de 10 a 100 vezes mais rápida do que a leitura de outras fontes de dados.

Outro fator importante e decisivo na utilização deste formato de arquivos para gravar as informações da base de dados, é a facilidade em se trabalhar com as informações de maneira a inserir e deletar registros quando se faz necessário. Além da agilidade de não precisar criar toda uma estrutura própria de um banco de dados ou gravar as informações em planilhas de excel.

É possível ver na Tabela 4 a comparação da diferença de velocidade na leitura pelo *QlikView* entre base de dados salvas em planilhas do Excel para com as informações salvas em bases .QVD.

Tabela 4 – Comparativo de velocidade de leitura no QlikView.

Base	Linhas	Colunas	Formato	Tamanho Arquivo	Tempo Carga
Base fictícia alteatória	108 mil	644	Planilha Excel	143 mb	16 segundos
			Arquivo .QVD	28 mb	3 segundos
Base Tweets Projeto	343 mil	149	Planilha Excel	847 mb	87 segundos
			Arquivo .QVD	187 mb	4 segundos

4.1.2 Coleta de Dados

Após definir a estrutura a ser utilizada para guardar com segurança os dados coletados a fim de garantir que não seja corrompida nenhuma informação, é hora de utilizar o *QlikView* para realizar a coleta da base de dados a partir das requisições na *API* do *Twitter* já explicadas anteriormente.

Um ponto importante a ser observado nesta etapa, é a realização deste processo em diversos períodos e em dias distintos com o objetivo de abranger a maior quantidade de usuários possíveis. Este método foi utilizado pois se a coleta fosse feita nos mesmos horários diariamente, haveria uma grande chance de exclusão de opiniões de possíveis usuários que utilizam a rede social em horários diversos. Desta forma, é possível garantir uma melhor distribuição de sentimentos na base.

Todo este processo foi estruturado de forma incremental a fim de garantir que um mesmo Tweet não fosse capturado duas vezes , com exceção dos RETWEETS.

Apesar de um RETWEET ser uma informação "duplicada", é um dado extremamente importante para o sistema, pois é um mesmo sentimento (positivo , negativo ou neutro) expresso por dois ou mais usuários que compartilham a mesma ideologia. Neste caso, há um flag na base que sinaliza essas situações.

Uma informação a ser considerada para a construção da base é o aumento do volume de dados nos períodos de maior movimento nas redes sociais, onde eram geradas as discussões nos dias anteriores e posteriores aos debates eleitorais realizados pelas principais emissoras de televisão de rede aberta do país, conforme Tabela 5 e Figura 5.

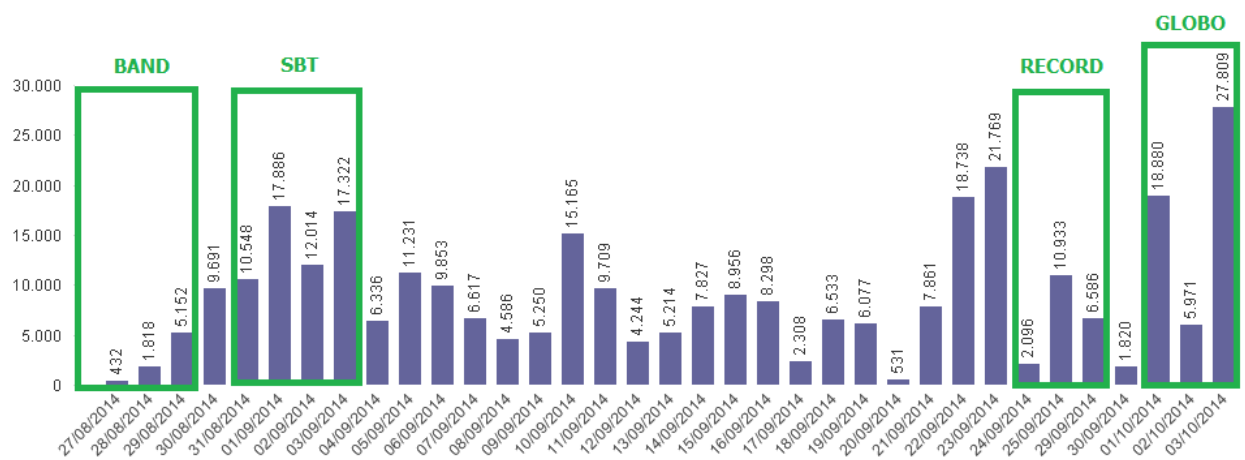


Figura 5 – Quantidade de registros coletados por data.

Tabela 5 – Datas dos Debates das principais emissoras de TV aberta.

Emissora de TV	Data do Debate
BAND	26/08/2014
SBT	01/09/2014
RECORD	28/09/2014
GLOBO	02/10/2014

4.2 Normalização

O maior objetivo desta etapa é fazer o melhor processamento possível na base de dados, a fim de garantir resultados mais incisivos durante as análises e comparações realizadas por este projeto. Como alguns dos usuários do *Twitter* utilizam de linguagem informal em suas publicações, pode ocorrer uma certa dificuldade no momento da realização de Classificação pelos algoritmos utilizados.

O foco deste trabalho é conseguir realizar este tipo de tratativa de normalização da forma mais automática possível, deixando de utilizar qualquer ação de tratativa manual devido ao grande volume de registros da base de dados.

4.2.1 Remoção de Links

Quando um usuário publica alguma informação através de um *tweet*, é possível que o mesmo consiga inserir links juntos ao texto, seja de uma matéria que ele gostou de ler, um vídeo ou qualquer outro tipo de informação que irá poluir a base de dados. Sendo assim, estes links devem ser removidos para não prejudicar a classificação.

Após o processamento desta etapa na base de dados, o resultado será conforme o demonstrado na Figura 6, com alguns exemplos de *tweets* pré e pós processamento:

Tweets - Original	Tweets - Sem Links
'Dilma deveria renunciar', diz Everaldo ao comentar denúncias na Petrobras - http://t.co/kPyPHNk9tj	'Dilma deveria renunciar', diz Everaldo ao comentar denúncias na Petrobras -
'É um vazio absoluto', diz ex-ministro Ciro Gomes sobre Marina Silva http://t.co/0rulOyaOb2	'É um vazio absoluto', diz ex-ministro Ciro Gomes sobre Marina Silva
'QUERO INTERNET DE ALTA VELOCIDADE NO BRASIL INTEIRO' http://t.co/c54aTgnbqv #Dilma13DeNovo	'QUERO INTERNET DE ALTA VELOCIDADE NO BRASIL INTEIRO' #Dilma13DeNovo
" Se houver lei no país, TSE deve cassar candidatura de Dilma Rousseff" http://t.co/xA2Zytaw4s	" Se houver lei no país, TSE deve cassar candidatura de Dilma Rousseff"
RT @jcmonteiro50: "@SakaSakamori: Desde morte Dudu, sou AÉCIO 45 http://t.co/baEGMRrENx	RT @jcmonteiro50: "@SakaSakamori: Desde morte Dudu, sou AÉCIO 45

Figura 6 – Tweets antes e após processamento de remoção de links.

O resultado deste processo é uma base otimizada que melhora o desempenho do algoritmo classificador.

Vale ressaltar também que um dos maiores benefícios é a remoção de quebras de linhas que podem existir no momento da extração dos dados do *Twitter*, as quais

atrapalham todo o processo, principalmente quando os dados são enviados para o Excel. Além disso essas quebras também dificultam a visualização dos dados.

4.2.2 Duplicados e ReTweet

O ReTweet é um função implementada pela rede social em questão para que os usuários possam compartilhar os sentimentos expressados por outro usuário, ou seja, é uma opinião compartilhada por dois ou mais usuários.

Geralmente este recurso é utilizado quando uma pessoa disserta sobre algum tema ou assunto de importância e que deve ser passado para frente por outros usuários a fim de alcançar um grande público e disseminar tal pensamento.

Ao buscar as informações da *API*, é possível identificar se um *tweet* foi escrito por aquele usuário ou se ele é um ReTweet. Tal informação é de grande importância para o sistema, pois assim esse ReTweet não será descartado, tendo em vista que é uma mesma opinião compartilhada por várias pessoas.

É preciso atentar para tal informação, pois o sistema foi desenvolvido para eliminar *tweets* duplicados, ou seja, a diferença de um *tweet* duplicado para um ReTweet é que no primeiro caso é um mesma informação propagada várias vezes pelo mesmo usuário, e não por usuários diferentes como no segundo caso.

Em suma, a opinião de várias pessoas sobre a mesma frase é importante e deve ser levada em conta, já a mesma opinião de apenas uma pessoa publicada por cem vezes, deve ter somente a importância do sentimento para apenas um usuário.

4.3 Classificação

Foram submetidos para a etapa de classificação de sentimentos todos os 330.881 registros coletados, conforme a distribuição informada na Tabela 2. Cada um dos grupos divididos por cada candidato teve seu processamento feito de forma individual através de chamadas da *API* já explicada anteriormente.

Este processo foi feito de forma manual, fora do *Qlikview*, porém para um desenvolvimento futuro e para melhor aproveitamento de todo o escopo deste projeto, uma integração entre todas essas etapas manuais é perfeitamente possível.

No total, foram gastos uma média de 1 a 2 minutos para o processamento de cada bloco de 10.000 *tweets*, com até 12 blocos por candidato, conforme limitações da *API* já explicadas anteriormente. É possível então considerar um resultado relativamente rápido, considerando o alto volume de dados que foram submetidos de uma só vez. Em possíveis aplicações futuras de outros projetos, o ideal é sempre realizar a carga incremental para

Em um cenário real de produção, tais gráficos podem ser atualizados facilmente pela aplicação de hora em hora por exemplo, em forma web facilmente acessados pelos usuários finais, através da própria plataforma disponibilizada pela aplicação.

4.4.1 Separação por Sentimentos

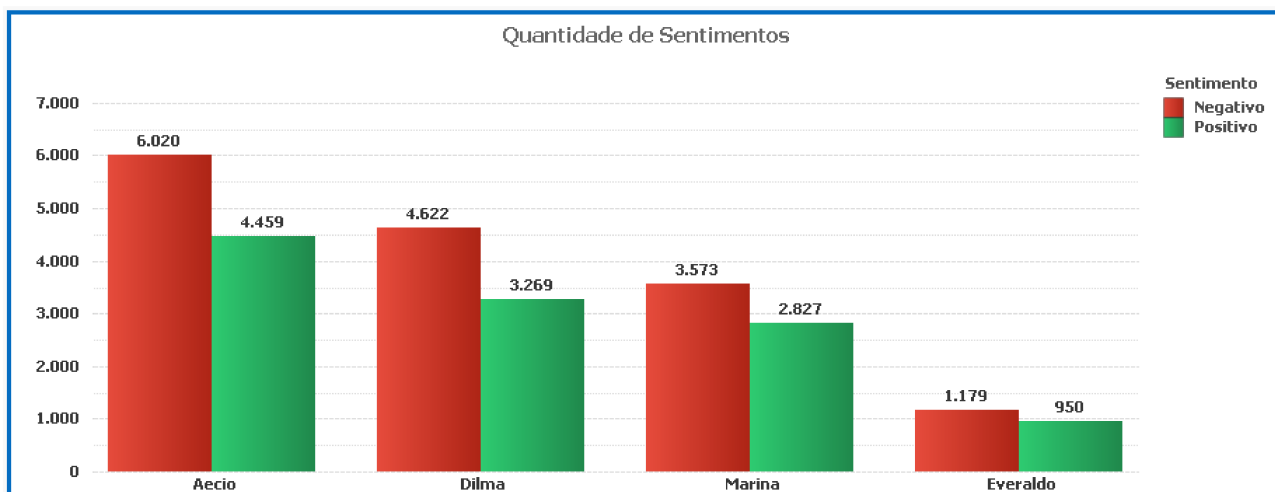


Figura 9 – Classificação por Sentimentos de cada Candidato.

A Figura 9 representa a separação da quantidade de sentimentos coletado por cada candidato pesquisado. O resultado final é a soma dos sentimentos alvo para este projeto, positivos e negativos, já que os sentimentos classificados como neutros não são mais importantes nesta etapa.

Ao analisar este gráfico, é possível verificar que para todos os candidatos, a maioria dos *tweets* foram de sentimentos negativos, ou seja, a maioria dos usuários utilizam a rede social em questão para expressarem o seu descontentamento aos políticos ao invés de mostrarem seu apoio aos mesmos.

É possível fazer um comparativo a um determinado produto que um consumidor compra em uma loja. A maioria das vezes em que ele gosta muito deste produto, ele não expressa este sentimento em uma rede social, mais caso este produto apresente algum tipo de defeito que faça o consumidor ficar insatisfeito, imediatamente ele acessa seu perfil para fazer uma determinada crítica sobre sua insatisfação.

4.4.2 Percentual de Classificados

Uma das principais dificuldades deste projeto é extrair do algoritmo classificador a maior quantidade possível de resultados satisfatórios na classificação de positivos e negativos. Tendo em vista este fato, existem várias etapas em que é possível considerar

que há a perda da qualidade dos sentimentos dos *tweets*, como já listadas, a normalização, a tradução da base e além do fato da própria curva de erro do algoritmo utilizado.

No gráfico de barras da Figura 10, estão expressos os valores exatos da soma de positivos e negativos em relação a base total que foi coletada. O resultado é um total de 26.899 *tweets* que efetivamente tiveram um sentimento identificado e classificado pelo algoritmo.

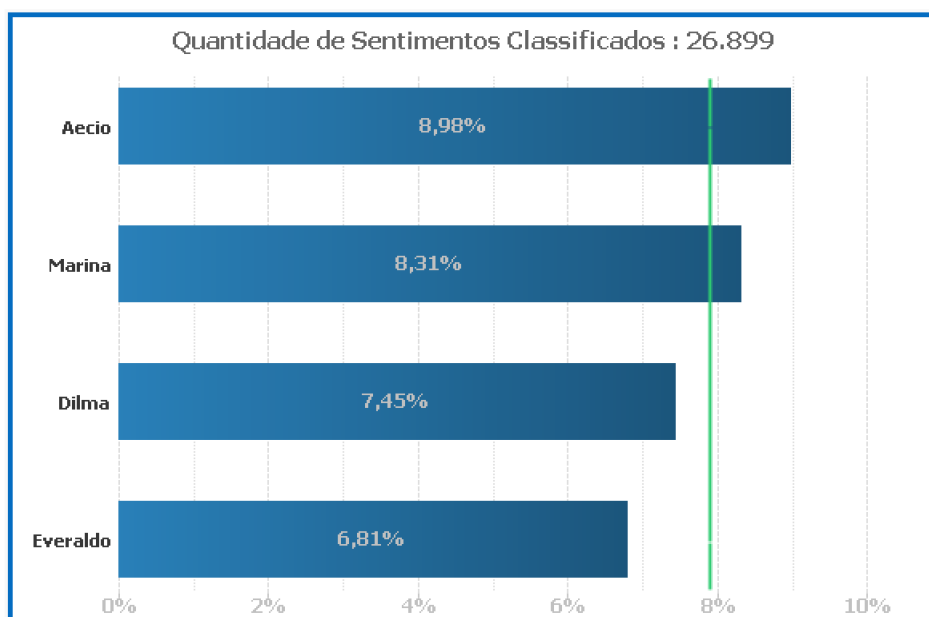


Figura 10 – Percentual de Tweets que foram Classificados.

Ao considerar todos estes fatores que influenciam a classificação de sentimentos, o resultado é um aproveitamento de cerca de 8,57% em média dos candidatos quanto a sua classificação, sendo que 91,43% foram classificados como neutro.

Dentro deste grande volume de neutros é possível considerar vários fatores, dentre eles, falha no algoritmo, dificuldade na interpretação e também na maioria dos casos não havia nenhum sentimento expresso no *tweet*.

Na Figura 11 estão listados alguns textos capturados que realmente não possuem nenhum sentimento envolvido, ou seja, são apenas publicações que foram descartadas por não representarem nenhum dado significativo de intenção de voto dos usuários.

ID	Classificação - Sentimento	Texto - Original	Candidato
508508378495483904	Neutro	Aécio Neves me processou vejam: http://t.co/kymKy4i3m8	
513756697408831490	Neutro	Mano Brown é Dilma, Chitaozinho é Aécio. Simbólico, eu acho.	
513757002472771585	Neutro	Como Malafaia e seus amigos estão reagindo ao crescimento de Dilma nas pesquisas... http://t.co/47CbnO8lU8	
513756845043752960	Neutro	"Com sua cara de vilã da Disney, Dilma Mocreff, na sua juventude, participou da guerrilha armada exercendo a função de canhão."	
513756697408831490	Neutro	Mano Brown é Dilma, Chitaozinho é Aécio. Simbólico, eu acho.	
513756398619230208	Neutro	Essa foto da Dilma hahahaha pqp	
513755927665594368	Neutro	'Forbes' lista cinco motivos para o Brasil não reeleger Dilma http://t.co/ApZbT2LwP0 via @VEJA	
513755400471318528	Neutro	A Dilma podia proibir as marcas de colocarem propagandas com a música do Pharrel em território brasileiro.	
513755209320108032	Neutro	RT @biosbug: Dilma lá e Padilha aqui. http://t.co/D9xt82JfF2	
514262045907099648	Neutro	Edir Macedo doa R\$ 10 milhões ao PT e manda Record atacar Marina Silva em seus telejornais.	
514261584932130816	Neutro	RT @gudiii: Marina Silva caiu de 40 para 26% das intenções de voto na região sul. Amém, meu povo acordou!	

Figura 11 – Resultados de classificações Neutras.

4.4.3 Intenção de Votos

Para a análise de intenção de votos dos *tweets* coletados, são considerados como universo total somente a soma dos sentimentos positivos, o qual servem para o propósito neste momento. Os negativos poderiam ser utilizados para a verificação da rejeição de cada candidato, ou seja, seria o resultado de uma pesquisa de quais candidatos a população jamais votaria em uma eleição.

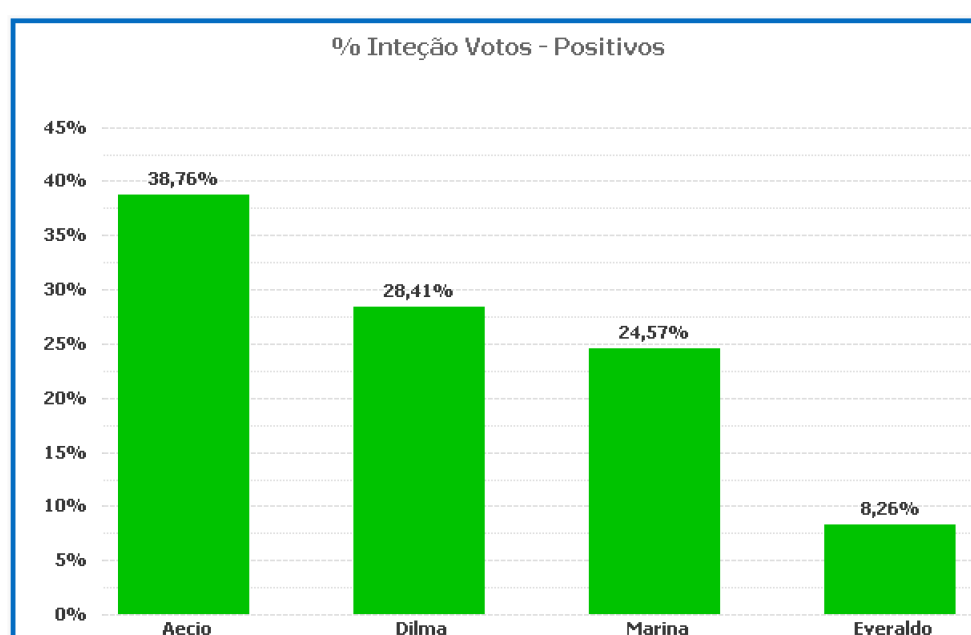


Figura 12 – Intenção de votos por Candidato.

Portanto, para tornar os resultados mais objetivos, somente os *tweets* que efetivamente expressam a intenção de voto de cada usuário foram somados para o resultado do gráfico abaixo. Sua fórmula é bastante simples, sendo composta pela divisão da soma dos *tweets* positivos de cada candidato pelo total de *tweets* positivos de todos os candidatos.

A Figura 13 revela a tabela oficial disponibilizada pelo site (TSE,) a qual demonstra a quantidade de votos válidos recebidos por cada candidato na eleição presidencial de 2014 no Brasil. Nela existe a representação da porcentagem de votos válidos recebidos por cada um dos candidatos.

Eleição: Eleições Gerais 2014 - 1º Turno - 05/10/2014 Abrangência: Brasil - Situação: Todas - Cargo: Presidente									
Brasil									
Última totalização: 22/10/2014 - 20:39:17									
UF	Cargo	Nr	Candidato	Partido	Coligação	Situação	Votação	% Válidos	
BR	Presidente	13	DILMA VANA ROUSSEFF	PT	PT / PMDB / PSD / PP / PR / PROS / PDT / PC do B / PRB	2º turno	43.267.668	41,59	
		45	AÉCIO NEVES DA CUNHA	PSDB	PSDB / PMN / SD / DEM / PEN / PTN / PTB / PTC / PT do B	2º turno	34.897.211	33,55	
		40	MARIA OSMARINA MARINA DA SILVA VAZ DE LIMA	PSB	PHS / PRP / PPS / PPL / PSB / PSL	Não eleito	22.176.619	21,32	
		50	LUCIANA KREBS GENRO	PSOL	PSOL	Não eleito	1.612.186	1,55	
		20	EVERALDO DIAS PEREIRA	PSC	PSC	Não eleito	780.513	0,75	
		43	EDUARDO JORGE MARTINS ALVES SOBRINHO	PV	PV	Não eleito	630.099	0,61	
		28	JOSÉ LEVY FIDELIX DA CRUZ	PRTB	PRTB	Não eleito	446.878	0,43	
		16	JOSÉ MARIA DE ALMEIDA	PSTU	PSTU	Não eleito	91.209	0,09	
		27	JOSE MARIA EYMAEL	PSDC	PSDC	Não eleito	61.250	0,06	
		21	MAURO LUÍS IASI	PCB	PCB	Não eleito	47.845	0,05	
		29	RUI COSTA PIMENTA	PCO	PCO	Não eleito	12.324	0,01	
Subtotal							104.023.802		
Subtotal							104.023.802		
Subtotal							104.023.802		
Total Geral							104.023.802		
Exportar dados									
1 - 11									

1 - 11

Figura 13 – Resultado da Eleição Presidencial no Brasil de 2014.

Na tabela 6 estão demonstradas as análises entre os resultados oficiais e a predição do projeto.

Tabela 6 – Comparativo de Intenção de Votos.

Candidato	Predição QlikView %	Resultado Eleição %
Dilma Rousseff	28,41	41,59
Aécio Neves	38,76	33,55
Marina Silva	24,57	21,32
Pastor Everaldo	8,26	0,75

Ao analisar todos os dados apresentados, é possível verificar que houve um aproximação bastante aceitável dos resultados obtidos durante este projeto, sendo em alguns casos bem parecidos com os das eleições.

A principal discrepância encontrada no comparativo da Tabela 6 foi em relação a candidata Dilma Rousseff, onde houve uma diferença significativa de 13,18 % desfavoráveis à predição do projeto. Segundo os dados já listados do (TSE,) a candidata em questão teve vitória por grande maioria dos votos exatamente nas regiões Norte e Nordeste.

Segundo censo do ano de 2010 sobre a inclusão digital no país e disponibilizado pela Fundação Getúlio Vargas (FGV) em (MAPAINCLUSAODIGITAL, 2010), o Brasil ocupa a 66ª posição no ranking global, com apenas 33% da sua população tendo acesso à Internet, bem atrás de países desenvolvidos como Suécia, Irlanda, Dinamarca e Estados Unidos que possuem índices superiores a 85%.

A pesquisa ainda disponibiliza um mapa (Figura 14) que avalia a quantidade de pessoas que possuem acesso a computadores com internet no Brasil.

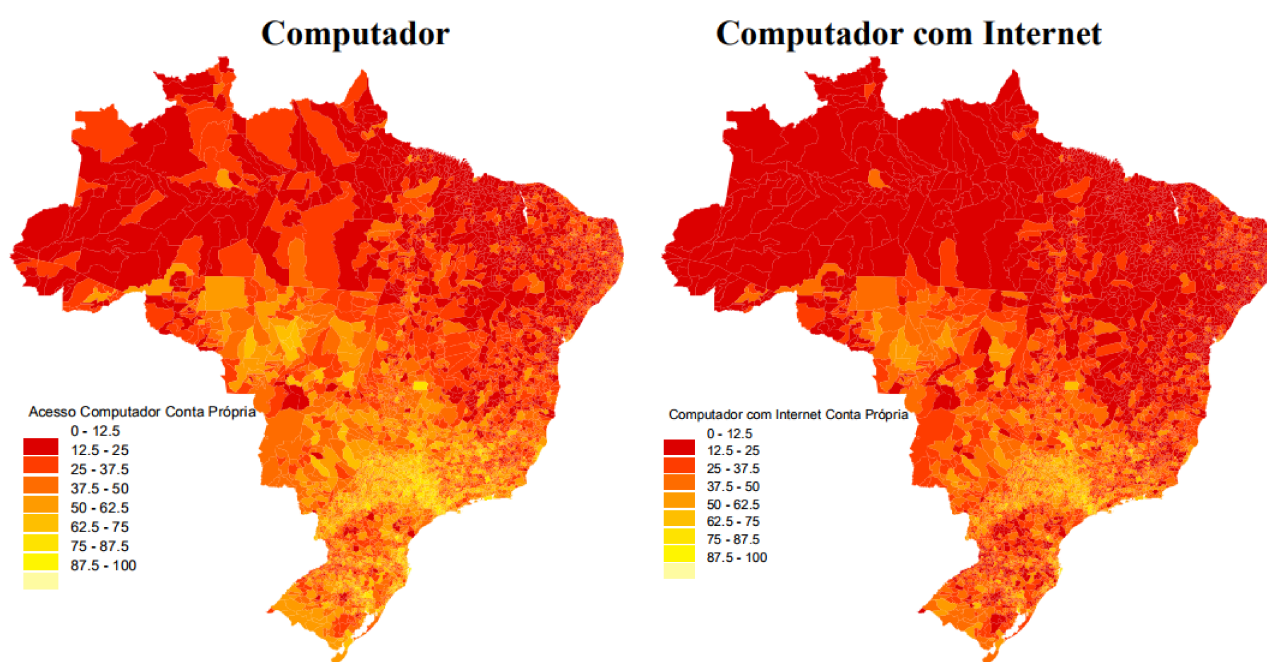


Figura 14 – Acesso a Internet conforme Censo 2010.

É possível perceber claramente no mapa que os principais pontos de baixo acesso a computadores com internet estão localizados nas regiões Norte e Nordeste.

Outro dado importante é que no perfil de cada usuário existe uma informação opcional que é a localidade em que o mesmo reside (Tabela 7). Esta tabela confirma a informação extraída da pesquisa da FGV, em que apenas 23% dos *tweets* identificados pertencem à região Norte e Nordeste.

Tabela 7 – Distribuição de localidade por região.

Região	Quantidade Tweets	% Região
Norte	4.468	4%
Nordeste	19.914	19%
Centro-Oeste	4.548	10%
Sudeste	57.885	54%
Sul	14.441	13%

Com base nestas duas informações extraídas, é possível especular e justificar a margem de erro obtida de 13,18 % na predição dos resultados para a candidata Dilma Rousseff, pois a maioria do seu possível público eleitor não é usuário da rede social *Twitter*.

5 Considerações Finais

5.1 Análise geral do trabalho

Vivemos em uma era de em que tudo se transforma cada vez mais rápido, novas formas de agir e pensar estão em constante modificação sempre para tornar as nossas vidas mais práticas e cómodas. Assim é em uma pesquisa eleitoral tradicional feita por grandes institutos de pesquisas como o IBOPE e o DATAFOLHA.

Em períodos de eleições milhares de funcionários dessas empresas tem o mesmo trabalho repetitivo de buscar a opinião de milhões de brasileiros ao longo do vasto território nacional. Muitas das vezes essas pesquisas são repetidas várias vezes durante o período eleitoral a fim de tentar medir quais são as intenções de votos da população Brasileira. Este projeto é facilmente expandido para eleições de todos os tipos como por exemplo, vereadores, deputados, senadores, prefeitos, governadores e todos os outros tipos de cargos públicos durante, antes ou depois das eleições municipais, estaduais e federais.

Atualmente várias grandes marcas e empresas fazem o monitoramento em tempo real de seus produtos e serviços nas redes sociais para saber qual o sentimento dos usuários em relação à elas, até mesmo programas de televisão tem utilizado este recurso, porém tal metodologia não é bem desenvolvida para o meio político atualmente.

Infelizmente conforme visto no final do Capítulo 4, um fator externo que não pode ser alterado e que afeta diretamente nos resultados das predições é o acesso da população a internet. Quanto maior a quantidade de usuários, melhores serão os resultados obtidos.

É claro que ainda há muito a ser desenvolvido nesta área de pesquisa e várias dificuldades foram encontradas neste projeto, porém os resultados são empolgantes e promissores.

5.2 Trabalhos Futuros

Este trabalho abre inúmeras possibilidades de modelos a serem desenvolvidos e melhorados para futuros processos de predição de resultados eleitorais. Abaixo estão listadas possíveis melhorias e ideias que podem ser implementadas para a busca de resultados cada vez mais próximos à realidade do cenário político atual:

- Aplicação em tempo real a nível de produção em que seriam avaliados resultados de hora em hora não só durante as campanhas, mais também antes e depois. Todas as

integrações que foram apresentadas neste trabalho funcionam perfeitamente entre si e todo o processo em que algumas etapas foram feitas de forma manual, podem ser implementadas para serem realizadas de forma automática. As aplicações no ramo da política podem ser várias, como:

- Aplicações para predições de eleições desde vereadores até presidentes.
 - Acompanhamento em tempo real da reação dos usuários das redes sociais após cada discurso feito pelo candidato, saber qual o seu desempenho em debates e quais os pontos que podem ser melhorados em sua campanha.
 - O candidato eleito pode acompanhar como esta a aceitação do seu mandato perante o sentimento dos usuários.
- Desenvolvimento de um algoritmo classificador já preparado para a linguagem portuguesa e especializado em assuntos políticos. Por melhor que seja a tradução utilizada, infelizmente os sentimentos podem ser perdidos durante o processo, por isso caso exista uma ferramenta apropriada na linguagem portuguesa, os resultados podem ser mais eficientes. É possível também a realização de um treinamento para os algoritmos baseados em aprendizado de máquina para tratarem assuntos políticos.
 - É comum nas redes sociais existirem muitas pessoas que são formadoras de opiniões e que possuem milhões de seguidores. É passível de um futuro desenvolvimento um modelo probabilístico com base no peso que cada usuário tem em uma rede social. Por exemplo, um *tweet* de algum ator muito famoso ou um astro da música sertaneja teria valor sentimental maior do que se comparado a um usuário que possui apenas dez amigos em sua lista, pois como ele é um formador de opiniões, várias pessoas podem mudar seu votos pela influência sofrida. Este é um ponto interessante que deve ser estudado e possivelmente implementado de forma muito bem balanceada a fim de que as predições não sejam afetadas negativamente.
 - Para a análise final deste projeto as classificações de *tweets* negativos não foram levadas em conta, pois o que importava era a intenção de voto. Porém o sentimento negativo pode ser utilizado para avaliar como esta a rejeição a um determinado candidato ou mandato, como por exemplo, em grandes escândalos políticos. Outras possibilidades são aplicações que monitoram os sentimentos dos usuários das redes sociais para a tomada de decisões políticas a qual são necessárias o apoio popular, como por exemplo plebiscitos e demais consultas públicas.

Referências

- ALGORITMOS. 2014. Disponível em: <<http://www.din.uem.br/~gpea/linhas-de-pesquisa/mineracao-de-dados/pre-processamento/pre-processamento-em-data-mining/>>. Acesso em: 20 mai. 2017. Citado na página 17.
- A.SHAMMA, D.; KENNEDY, L.; CHURCHILL, E. F. Tweet the debates. *EE Columbia*, n. 1, 2012. Citado na página 12.
- BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, n. 1, 2011. Citado na página 15.
- BIFET, A.; FRANK, E. Sentiment knowledge discovery in twitter streaming data. *University of Waikato, Hamilton, New Zealand*, n. 1, 2010. Citado na página 16.
- BOTELHO, F. G.; UGIONI, P. H. R. *Data Mining em Redes Sociais*. Tese (Trabalho de Conclusão de Curso) — Universidade Federal de Santa Catarina, Florianópolis - SC, 2012. Citado 2 vezes nas páginas 13 e 16.
- BRASILDIGITALFUTURE. Disponível em: <<https://www.comscore.com/por/Insights/Presentations-and-Whitepapers/2014/2014-Brazil-Digital-Future-in-Focus-Webinar>>. Acesso em: 22 ago. 2014. Citado na página 11.
- CEVI, C. R. *Um Estudo sobre Mineração de Dados em Redes Sociais*. Tese (Trabalho Individual II) — Universidade Federal do Rio Grande do Sul, Porto Alegre - RS, dez. 2008. Citado na página 11.
- CHEN, H.; ZIMBRA, D. Ai and opinion mining. *IEEE Intelligent Systems*, v. 1, n. 1541, p. 74-80, 2011. Citado na página 11.
- FOLHASAOPAULO. 2017. Disponível em: <<http://www1.folha.uol.com.br/tec/2017/02/1861175-numero-de-usuarios-do-twitter-no-brasil-cresce-18-em-2016.shtml>>. Acesso em: 24 jul. 2017. Citado na página 11.
- GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *Sentiment140.com*, n. 1, 2013. Citado 2 vezes nas páginas 20 e 30.
- GOOGLETRANSLATE. 2016. Disponível em: <<https://translate.google.com.br/?hl=pt-BR>>. Acesso em: 22 mai. 2016. Citado na página 29.
- KRAEMER, A.; KAWAMOTO, A. L. S.; GEROSA, M. A. *Predição de Resultado de eleição para Reitor de Universidade usando Tweets como fonte de pesquisa*. Tese (Doutorado) — Universidade Federal Tecnológica do Paraná e Universidade de São Paulo - USP, Paraná / São Paulo, 2012. Citado 2 vezes nas páginas 12 e 20.
- LIU, B.; ZHANG, L. *A Survey of opinion mining and Sentiment Analysis*. Tese (Doutorado) — University of Illinois at Chicago, Chicago, IL, 2012. Citado na página 18.

- MAPAINCLUSAODIGITAL. 2010. Disponível em: <<http://www.cps.fgv.br/cps/telefonica/>>. Acesso em: 28 jul. 2017. Citado na página 42.
- MICROSOFTARTICLES. 2017. Disponível em: <<http://msdn.microsoft.com/pt-br/library/ms174949.aspx>>. Acesso em: 12 jul. 2017. Citado na página 17.
- QLIK. 2017. Disponível em: <<http://www.qlik.com/us/products/qlikview>>. Acesso em: 24 mai. 2015. Citado na página 33.
- RESTAPIS. 2014. Disponível em: <<https://dev.twitter.com/rest/public>>. Acesso em: 30 oct. 2014. Citado 2 vezes nas páginas 23 e 26.
- SENTIMENT140. 2017. Disponível em: <<http://www.sentiment140.com/>>. Acesso em: 27 jul. 2017. Citado na página 30.
- SOUZA, D. *Marketing Digital - Mídias Sociais*. Tese (Trabalho de Conclusão de Curso) — CEETEPS - Centro Estadual de Educação Tecnológica Paula Souza, Taubaté - SP, 2013. Citado na página 15.
- TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. *Revista Ibérica de Sistemas e Tecnologias de Informação*, v. 1, n. 8, p. 53–65, 2011. Citado 2 vezes nas páginas 11 e 19.
- TIPOSAPIS. 2016. Disponível em: <<http://www.linhadecodigo.com.br/artigo/3471/utilizando-a-api-do-twitter-no-desenvolvimento-de-aplicacoes-web-com-php-e-curl.aspx>>. Acesso em: 22 mai. 2016. Citado na página 23.
- TODAMATERIA. 2017. Disponível em: <<https://www.todamateria.com.br/linguagem-coloquial/>>. Acesso em: 19 jul. 2017. Citado na página 28.
- TSE. Disponível em: <<http://www.tse.jus.br/eleicoes/estatisticas/estatisticas-candidaturas-2014/estatisticas-eleitorais-2014-resultados>>. Acesso em: 26 jul. 2017. Citado 2 vezes nas páginas 41 e 42.
- TUMASJAN, A.; SPRENGER, P. G. S. T. O.; WELPE, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Fourth International AAAI Conference on Weblogs and Social Media*, n. 1, 2010. Citado na página 19.
- VILELA, P. de C. S. *Classificação de Sentimento para Notícias sobre a Petrobras no Mercado Financeiro*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro - PUC-RJ, Rio de Janeiro, 2011. Citado na página 18.