Aplicação para Identificar a Opinião de Eleitores Brasileiros, tendo como Estudo de Caso, Deputado Jair Messias Bolsonaro

André Felipe Gouveia Farias[1], Vinícius de Oliveira Andrade[2,] Ellen Souza[1]

Curso de Bacharelado em Sistemas de Informação – Universidade Federal Rural de Pernambuco (UFRPE) – Campus Serra Talhada – PE– Brasil

andrefelipegf@gmail.com, viniciusibsm@gmail.com, eprs@uast.ufrpe.br

Abstract. *In the current era with a technological globalization, different communications networks are emerging every day, allowing different people from all over the world to interact, generating massive data values. Based on the above, the objective of the present work is to implement text mining techniques in order to know how the opinions of the principals regarding the presidential candidate of Brazil, Deputy Jair Messias Bolsonaro, through opinion mining of voters who use the social network called Facebook. The accuracy of 82.60% was obtained with the balanced data set with two classes (positive and negative), using the SVM algorithm.*

Resumo. *Na época atual com uma globalização tecnológica, vêm surgindo a cada dia diferentes redes de comunicações, permitindo que diferentes pessoas de vários lugares do mundo interajam, gerando maciças quantidades de Dados. Com base no que foi citado, o objetivo do presente trabalho, é implementar técnicas de mineração de texto com o intuito de saber as opiniões dos eleitores em relação ao candidato a presidência do Brasil, Deputado Jair Messias Bolsonaro, por meio da mineração de opinião de eleitores que usam a rede social chamada Facebook. A acuráacia de 82,60% foi obtido com o conjunto de dados balanceados com duas classes (positivo e negativo), utilizando o algoritmo SVM.*

.

1. Introdução

No momento, um dos temas mais discutidos é sobre política, pois estamos muito próximos de mais uma eleição presidencial, essas discussões estão por toda parte e um lugar onde o tema é bastante argumentado é nas redes sociais como, por exemplo, o Facebook com “117 milhões” (Cruz, 2017) de usuários

brasileiros ativos mensalmente, discutindo os mais diversos assuntos como: política, saúde, segurança e etc.

No atual momento, o Facebook é a rede social mais utilizada no mundo, gerando dois bilhões de usuários ativos por mês (Cruz, 2017), o Brasil como já citado representa uma parcela significativa desses usuários (117 milhões) e estar em primeiro lugar como o país que mais cresce na rede social citada (Socialbakers, 2011). Levando em consideração o que já foi dito, podemos considerar, o Facebook como uma grande rede de dados de informações sendo um lugar propício para a mineração de opiniões.

Mineração de opinião, termo que já citado, seria o tratamento de um texto onde o principal objetivo seria extrair a opinião do usuário sobre um determinado tema, essas opiniões podem ser positivas, negativas, ambas, irrelevante e irônica.

O presente artigo tem como objetivo, minerar opiniões sobre o candidato a presidência do Brasil, Deputado Jair Messias Bolsonaro e a mineração dessas opiniões foram feitas na rede social chamada Facebook. Foram aplicadas técnicas de mineração de opinião, sobre as opiniões dos usuários do Facebook, com o intuito de avaliar cada uma dessas opiniões sobre o Deputado Jair Messias Bolsonaro.

A seguir, na seção 2, é apresentada uma pesquisa de alguns trabalhos relacionados, usando mineração de opinião, na seção 3 é apresentada os métodos aplicados para o objetivo proposto, na sessão 4, é apresentado os resultados obtidos conforme as técnicas utilizada e apresentada na seção 3. Em seguida a conclusão do presente trabalho.

2. Trabalhos Relacionados

“Mineração da Opinião ou Análise de Sentimento (AS) é o estudo computacional de opiniões, sentimentos e emoções expressas acerca de entidades, eventos e seus atributos, que estão em um texto.” (LIU, 2010). Essas opiniões são facilmente encontradas em redes sociais como, por exemplo, Facebook ou Tweeter. Podem também ser encontrar opiniões em sites de notícias ou bloggers.

“Mineração da Opinião é o problema de identificar opiniões expressadas sobre um determinado assunto e avaliar a polaridade dessa opinião.” (TSYTSARAU; PALPANAS, 2012). As opiniões são classificadas em categorias como: negativa, positiva, neutra, irrelevante (textos não relacionados à temática do projeto) ambas (textos subjetivos contendo palavras positivas e negativas) e irônica.

A geração das bases de dados pode ser através da API (Application Programming Interface) do Facebook, os comentários foram salvos em formatos JSON, onde cada linha representava uma opinião.

O trabalho de Thiago et. al (2013) analisa o sentimento acerca dos protestos que ocorreram no Brasil, onde foram coletados 300 mil Tweets. Os mesmo foram classificados entre apoio ou repúdio as manifestações em seguida passaram por um pré-processamento onde foram filtradas as mensagens, caracteres de pontuações, URLs, Stopwords. A última etapa do pré-processamento foi o processo de stemização. As métricas utilizadas foram a acerácea, a variância, o desvio, o desvio padrão, precisão, recall e Macro Avaraged. Os resultados obtidos pelo classificador Naives Bayes foram satisfatórios com uma precisão de 90%.

O trabalho de Dattu et al. (2015) é a comparação entre técnicas de analise de sentimento. As técnicas analisadas foram a TwitterSentiment, SentiStrength, SentiStrength+ TwitterSentiment, MNB, NBSVM, NB, SVM. Foram utilizadas bancos de dados do twitter sobre vários temas, com cada técnica. Após a analise dos resultados, foi constatado que o SVM teve uma acurácia maior , com 89,8%, seguida do NB com 89,4%. Foi constatado que Naive Bayes classificador é insensível a dados desequilibrados que dão resultados mais precisos.

3. Método

A Figura 3.1 Mostra as etapas que foram executadas para realizar a mineração de opinião. Cada etapa será descrita nas seções subseqüentes.

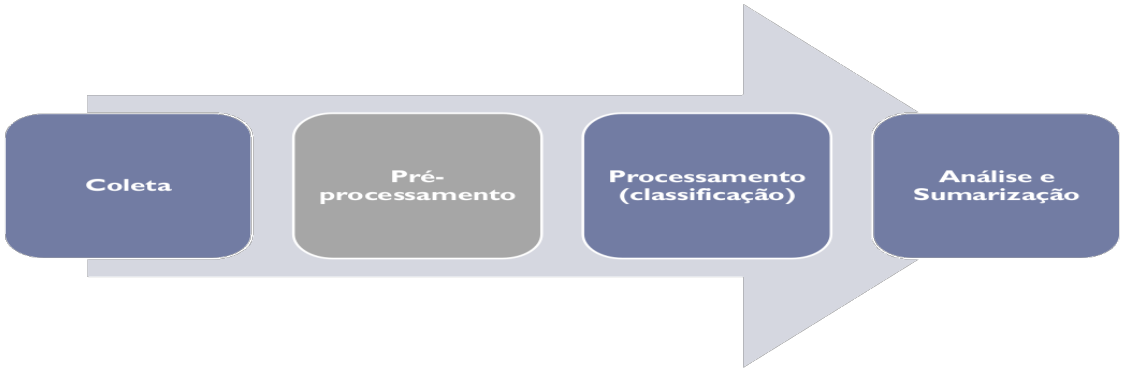

Figura 3.1 - Método proposto
Fonte: Elaborada pelo autor, 2017

3.1 Coleta dos Dados

A criação do Corpus foi realizada em três passos: extração dos comentários do Facebook, filtragem dos mesmos e anotação manual;

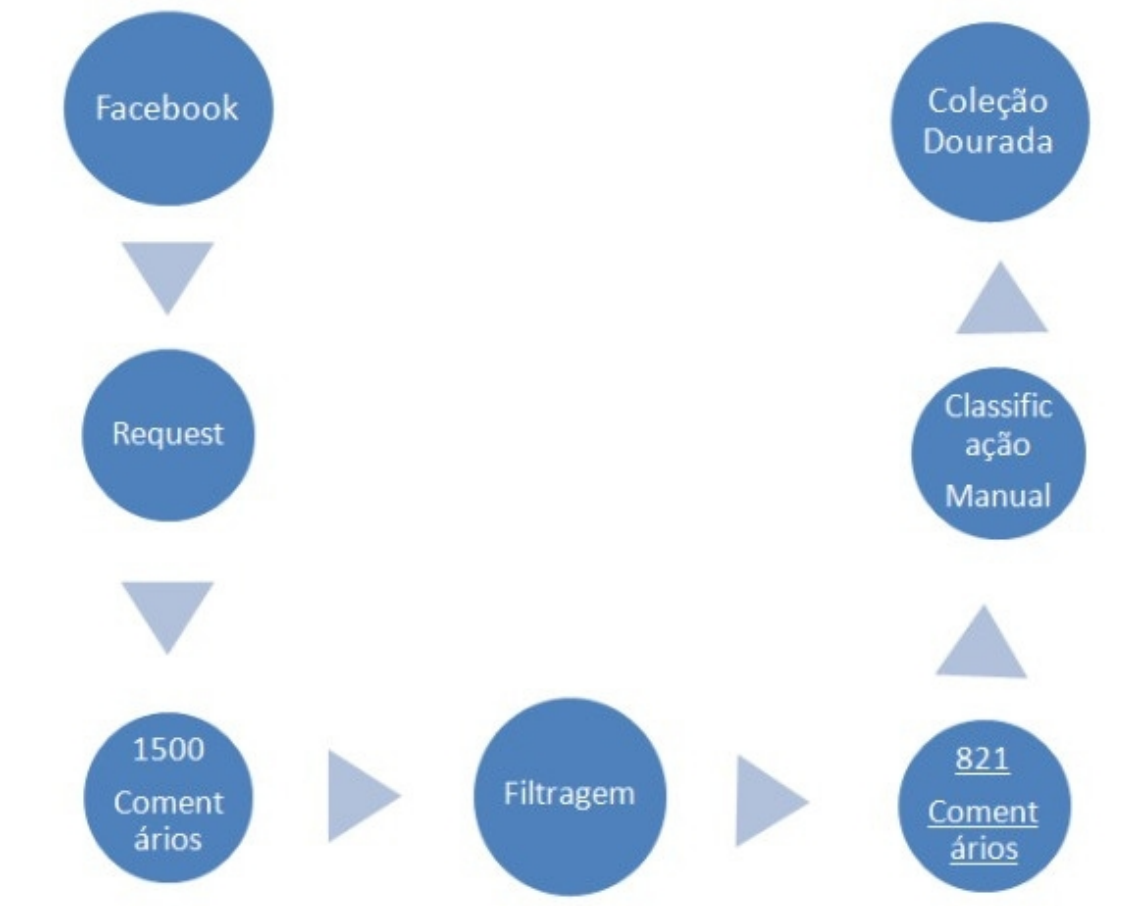

Figura 3.2 – Processo de Coleta dos Dados.
Fonte: Elaborada pelo autor, 2017

A coleta dos comentários deu início com o uso da API do Facebook e da biblioteca Request recebendo respostas em JSON. Inicialmente foram coletados 1500 comentários onde os mesmos foram filtrados, removendo-se 679 comentários do Facebook que poderiam conter apenas: hastags, emotions e citações. Os comentários removidos não exerceriam influencia na aprendizagem de máquina. Os 821 comentários do Facebook que restaram foram classificados, em seis classes: positivo, que são textos subjetivos contendo apenas palavras positivas, sobre o candidato, negativo, que são textos subjetivos contendo apenas palavras negativas sobre o candidato, ambas que são textos subjetivos contendo palavras positivas e negativas sobre o candidato, neutra que são textos objetivos sobre o candidato, ou seja, sem qualquer opinião, irrelevante que são textos não relacionados à temática do projeto, entre outros, ironia que são textos com opinião, mas que contém emoticon ou hastag que indica opinião contrária. com os respectivos

identificadores: 1, 2, 3, 4, 5 e 6. Para a coleção dourada final, foram utilizados apenas os comentários classificados como positivo e negativo.

A Tabela 3.1 ilustra como ficou dividida a coleção dourada, tanto desbalanceada quanto balanceada final.

Desbalanceada						
	Positivo	Negativo	Ambas	Neutra	Irrelevante	Ironia
Quantidade	414	201	9	122	41	34
Balanceada						
	Positivo	Negativo	Ambas	Neutra	Irrelevante	Ironia
Quantidade	201	201	0	0	0	0

Tabela 3.1 – Coleção Dourada dividida em 6 classes

3.2 Pré-processamento do Texto

Os textos expostos no Facebook são em geral curtos e informais escritos em uma linguagem coloquial, e em uma quantidade pequena de palavras, e de maneira desestruturadas. Os aplicativos de mineração de dados, em sua maioria, assimilam apenas textos estruturados. Para dar início aos métodos de aprendizagem precisam de textos estruturados.

Nessa etapa do processo, as técnicas para o tratamento dos comentários, foram tokenizing , remoção de caracteres especiais, remoção de stopwords, remoção de hashtags, remoção de URLs, remoção de nomes dos políticos e stemming, em seguida junto da sua polaridade, serão convertidos para que se torne possível aplicar o aprendizado de maquina.

A aplicação foi desenvolvida em Python3, para a etapa de pré-processamento foram utilizadas as bibliotecas Unicode Data, Regular Expression e NLTK as quais disponibilizam vários recursos para esse processo.

3.2.1. Tokenização

É o processo de quebra de texto em palavras ou sentenças, para isto foi utilizado o modulo da biblioteca NLTK, TweetTokinizer, por se encaixar melhor com o estilo de tokenização que era necessário para a pesquisa.

3.2.2. Remoção de Caracteres Especiais, Emotions, URL, Bolsonaro e Hashtags

Para esta tarefa foram utilizadas expressões regulares disponíveis no Python3 para a remoção dos artigos desnecessários que não agregam valor a opinião do usuário, assim provendo de um corpus limpo e livres de dados desnecessários

3.2.3. Remoção de stopwords

Existem palavras comuns, de pouco valor para a aplicação, consequentemente podem ser excluídos do vocabulário, sem atribuir dano algum ao trabalho. Foi utilizada para a remoção das palavras uma lista de Stopwords disponíveis na biblioteca NTKL do Python.

3.2.4. Stemming

Para esta tarefa foi aplicada a função de stemming da biblioteca NLTK para transformação das palavras do corpus em seus radicais.

3.3Processamento

Nesta etapa, é realizada a classificação da polaridade dos documentos obtidos através da coleta de dados feito no Facebook. Utilizando-se a biblioteca Scikit-Learn, pois tem código aberto de aprendizado de maquina, é desenvolvida na linguagem de programação Python.

Do Scikit-Learn os recursos utilizados foram: do Naıve-Bayes o algoritmo Multinomial, além do Countvectorizer, Tfidfvectorizer, Kfold, o algoritmo SVC Linear e o algoritmo de Regressão Logística.

4. Resultado

Os resultados obtidos do processamento, usando três tipos de configuração de pré-processamento, são eles: o Bag of Word com Frequência, o Bag of Word com TF-IDF e Bag of Word com Frequência e Word Bigram, estão dispostos na tabela 4.1.

Id	Pré-processamento	Naive Bayes	Support Vector Machine	Regressão Linear
1	Tokenização Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos Stemização	80,95%	**82,60%**	79,68%

	Bag-of-Word + Frequencia			
2	Tokenização Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos Stemização Bag-of-Word + TF-IDF	81,13%	80,65%	80,32%
3	Tokenização Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos Stemização Bag-of-Word + Frequencia + Word-Bigram	81,30%	80,49%	80,97%

Tabela 4.1 – Resultado da etapa de processamento

Para a avaliação dos resultados, foi utilizado o método de validação cruzada *k-fold*. Os resultados foram obtidos utilizando a medida de acurácia.

4.1. Validação Cruzada e K-Fold

A técnica de validação cruzada *k-fold* consiste na divisão da coleção dourada em *k* blocos de mesmo tamanho. As etapas de treinamento e teste são realizadas *k* vezes, onde é utilizado um dos blocos para teste e os demais para treinamento. A cada iteração, o bloco escolhido para teste muda. Para o presente trabalho foi utilizado k=4.

4.2. Acurácia

A acurácia é uma medida precisa e efetiva, é frequentemente utilizada para avaliar aprendizado de máquina.

Ao final de cada rodada do *K-fold* a acurácia é computada e no final é calculada a média. Para assim ser usada como base para os resultados. A Tabela 4.1 demostra os resultados para etapa de processamento utilizando a validação cruzada *K-fold*.

4.3. Matriz de confusão

É um tipo de tabela que permite a visualização do desempenho de um algoritmo de aprendizado. Cada coluna da matriz representa as instâncias de uma classe prevista, enquanto as linhas representam os casos de uma classe real.

As Tabelas 4.2, 4.3, 4.4 e 4.5, ilustram a matriz de confusão para cada rodada do K-fold do melhor resultado apresentado anteriormente.

		Previsão	
		Positivo	Negativo
Real	Positivo	37	9
	Negativo	12	43

Tabela 4.2 – Primeira rodada

		Previsão	
		Positivo	Negativo
Real	Positivo	41	11
	Negativo	9	40

Tabela 4.3 – Segunda rodada

		Previsão	
		Positivo	Negativo
Real	Positivo	45	13
	Negativo	8	34

Tabela 4.4 – Terceira rodada

		Previsão	
		Positivo	Negativo
Real	Positivo	40	5
	Negativo	3	52

Tabela 4.5 – Quarta rodada

Como pode ser observado na Tabela 4.1,o melhor resultado obtido foi utilizando o SVM na a configuração 1 (Tokenização, Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos, Stemização, Bag-of-Word + Frequencia), onde obteve uma acurácia de 82,60%.

5. Conclusão

Foi possível chegar a conclusão de que através da mineração de dados nas redes sociais, é possível afirmar a opinião de uma população sobre determinado candidato.

Podemos concluir sobre o Dep. Jair Messias Bolsonaro, é que ele é bastante citado na rede social e que na maioria das publicações referente a sua pessoa, são de usuários com opiniões positivas.

Aplicando mais técnicas de pré-processamento e processamento, poderia até saber sobre o padrão de classe social dos usuários, como poderia descobrir de qual estado da região do Brasil o usuário que fez o comentário mora. Ficando como proposta para trabalhos futuros.

6. Referências Bibliográficas

LIU, B. Sentiment Analysis and Subjectivity. **Handbook of Natural Language Processing**, n. 1, p. 1–38, 2010.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. **Data Mining and Knowledge Discovery**, v. 24, n. 3, p. 478–514, 2012.

CRU, Melissa. **Facebook chegou a 2 bilhões de usuários.** Disponível em: <https://www.techtudo.com.br/noticias/2017/06/facebook-chega-a-2-bilhoes-de-usuarios.ghtml>. Acesso em: 18 de fevereiro 2018.

Socialbakers. 10 Fastest Growing Countries on Facebook. Disponível em: <https://www.socialbakers.com/blog/1290-10-fastest-growing-countries-on-facebook-in-2012>. Acesso em: 18 de fevereiro 2018.

DATTU, B. S.; GORE, P. D. V. A Survey on Sentiment Analysis on Twitter Data Using Different Techniques. v. 6, n. 6, p. 5358–5362, 2015.