Analise de sentimentos expressos no twitter durante campanhas presidências vitoriosas

Abílio N. Barros[1], **Brenno L. Barbosa**[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE– Brazil

{abilio.nogueira,brenno.luiz}@ufrpe.br

Abstract. *The present work deals with an analysis of feelings based on the twitters of presidents Jair Messias Bolsonaro (Brazil) and Donald Jhon Trump (USA). Both candidates have used their social networks to promote their election campaigns. Throughout the article, we will observe the similarities in feelings expressed by them in the period of one year to a few months before their respective elections.*

Resumo. *O presente trabalho trata-se de uma análise de sentimentos, tendo como base twitters dos presidentes Jair Messias Bolsonaro (Brasil) e Donald Jhon Trump (EUA). Ambos os candidatos se utilizaram de suas redes sociais para promover suas campanhas eleitorais. Ao decorrer do artigo, iremos observar as semelhanças nos sentimentos expressos por eles no período de um ano até alguns meses antes de suas respectivas eleições.*

1. Introdução

O cenário político do Brasil nas eleições de 2018 deu um grande destaque e popularizou quem, em 2019, é presidente do Brasil, Jair Messias Bolsonaro. Com 55% dos votos, segundo o TSE em 2018, Jair teve como seu apoio nas eleições o uso das redes sociais, que contribuíram para sua popularização no país. Uma das comparações utilizadas na eleição de 2018, foi a de Jair Bolsonaro com Donald John Trump, presidente do Estado Unidos. Donald Trump, assim como Bolsonaro, se utilizou das redes para propagar seu discurso político, participando ativamente na internet, com perfis no Facebook, Twitter, entre outros.

O uso da redes ajudou ambos candidatos a chegar à presidência, e embora tenham conseguido seus objetivos, a comunicação com seus eleitores nas redes ainda não acabou. Segundo o professor Alberto Valle, da Academia do Marketing do Rio de Janeiro," a tecnologia também precisa ser utilizada para criar relacionamento com o público fora da época de eleições".

Com o conteúdo exposto, podemos observar semelhanças entre o presidente brasileiro e o estadunidense. O conteúdo descrito em suas redes sociais, neste trabalho em particular o Twitter, irá demonstrar se existe alguma simetria no posicionamento dos candidatos.

Com base de dados retirados dos twitters dos presidentes, foi feito uma análise de sentimentos buscando caracterizar as emoções expressas nas postagem em: positivo, negativo e neutro. Técnicas de aprendizagem de máquina foram utilizadas para realizar as análises de dados. Ao decorrer do trabalho, serão apresentados os processos utilizados para o experimento, tais como: materiais, métodos, prática e resultado.

2. Materiais

O ramo de aprendizagem de máquina é um sub-campo da inteligência artificial , cujo busca-se por meios de algoritmos ensinar aos computadores qual é o resultado daquela ação que aconteceu anteriormente, buscando assim, que ele consiga predizer o resultado de uma futura ação com características parecidas.

Dentre os vários tipos de aprendizado, nesse artigo, é abordado a ideia de aprendizagem supervisionada, no qual antes dos dados serem entregues para a análise da máquina esses dados são rotulados por humanos, para que assim, a máquina só precise entender como sair de um ponto A até um ponto B, tendo esses pontos já mapeado nos dados.

Existem mais de uma forma de analisarmos sentimentos em texto, no entanto, durante este trabalho utilizamos apenas métodos estatísticos desconsiderando assim o significado semântico de cada palavra, e com isso, é possível remover a problemática de usar duas línguas diferentes, já que os autores das mensagens utilizavam-se de seus idiomas nativos.

2.1. Naive Bayes

É um algoritmo classificador probabilístico baseado no teorema estatístico de Bayes, sendo popular na área de aprendizagem de máquina, pois, caracteriza aquela texto estudado pela frequência das palavras pertencentes ao texto.

Um ponto importante para sua escolha foi o fato dele desconsiderar eventos correlatos , como buscamos categorizar cada mensagem expressa sem a relação com as outras, tal algoritmo se mostra mais eficiente e requerendo menos poder computacional.

2.2. Python

Python é uma linguagem de alto nível que possui várias bibliotecas para análise de dados e processamento de quantidade de dados.

2.3. NLTK(Natural Language Toolkit)

Utilizamos também a plataforma NLTK que é um conjunto de ferramentas utilizados no python para processamento de linguagem natural, onde encontramos implementações do algoritmo utilizado e algumas métricas utilizadas para processamento de palavras pela comunidade.

2.4. Base de Dados

Neste trabalho foram utilizadas quatro bases de dados distintas:

Base para treinamento em português Uma base retirada do Site kaggle com twitters já rotulados no idioma português brasileiro.

Base para treinamento em inglês Uma base retirada de perfil de um usuario do Site kaggle com twitters já rotulados em inglês.

Base com twitters de Jair Bolsonaro Uma base retirada de perfil de um usuario do Site kaggle com twitters do presidente brasileiro.

Base com twitters de Donald Trump Uma base retirada de perfil de um usuario do GitHub com twitters do presidente norte americano.

2.5. Google Colab

Para criar dois classificadores, um para português e outro para inglês, e ter um aceleramento por parte da GPU dos computadores utilizados, o Google Colab mostrou-se a melhor ferramenta para o desenvolvimento do trabalho.

3. Métodos

3.1. Limpeza e colunas e alguns dados

Primeiro foi realizado uma limpeza nas colunas que compunham as base de dados, nas bases de treinamento e teste foram apenas as mensagens e os rótulos, e a das personalidades políticas, apenas a mensagem e as datas. Também foram removidos, quase que em sua totalidade, todos os emoticons. Já quanto às palavras maiúsculas e minúsculas todas foram reduzidas a minúsculas e no caso das bases em português-brasil foram removidos suas acentuações gráficas.

3.2. Balanceamento das Bases

As bases sofreram balanceamentos para seu uso neste trabalho, por não se tratarem da mesma fonte todas foram reduzidas.

Para a base de aprendizado na língua inglesa foram utilizado 2361 sentenças de cada rótulo (positivo,negativo e neutro) totalizando 7083 sentenças rotuladas. Já para a base em português brasileiro foram utilizadas 482 sentenças de cada rótulo totalizando 1446 frases. Cada base de aprendizado é dividida entre treinamento e teste, tomando as proporções de 70% e 30% respectivamente. Quanto às bases com tweets dos políticos foram reservados uma mesma faixa temporal alterando só o ano, já que as eleições dos Estados Unidos da América foram realizadas em 08 de novembro de 2016 e no brasil em 07 de outubro de 2018. Com isso, as mensagens selecionadas obedeceram as seguintes faixas temporais 01/08/2015-16/08/2016 e 01/08/2017-16/08/2018.

3.3. Removendo palavras sem valor

Nominadas de stopwords(palavras vazias,em tradução livre) são palavras que comuns que não empregam nenhum sentido a frase e com isso devem ser retiradas para que as tabelas construídas sejam menores.O nltk já oferece bases com palavras comuns para diversos idiomas.

3.4. Removendo radicais

A técnica denominada de stemming é a prática de reduzir as palavras a seu radical primitivo reduzindo ainda mais a quantidade de palavras a serem trabalhadas.

4. Experimentos

Por estarmos trabalhando com dois idiomas distinto foi visto a necessidade de criar dois classificadores distintos e realizar seu treinamento seguindo as as etapas descritas na seção anterior(Métodos). Após realizar o processo de treinamento dos classificadores , foi executado a etapa de medir sua precisão por meio da acurácia expressa pelos classificadores ao executarem os testes em suas respectivas bases de teste, com isso foi expresso os seguintes resultados:

Classificador em Inglês Obteve 67,64%
Classificador em Português Obteve 62,80%

4.1. Matriz de Confusão

A matriz de confusão , também chamada de tabela de confusão, é uma matriz composta pelas linhas sendo nosso resultado já rotulado e a coluna como o resultado impresso pelo classificador. Assim podemos ter um visão de onde nosso classificador pode estar errando.

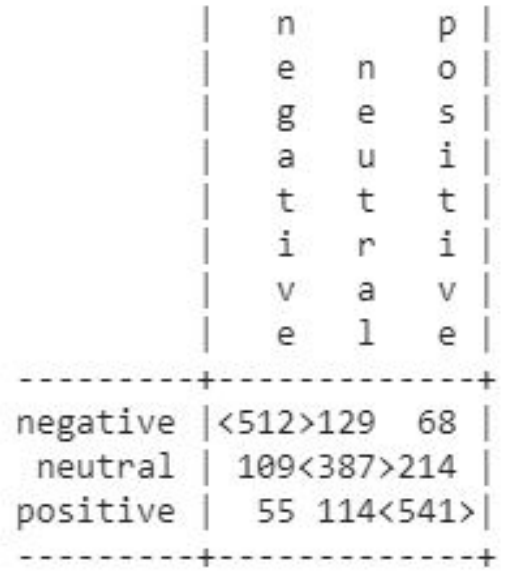

Figure 1. Matriz de Confusão do Classificador Inglês

Nessa matriz podemos ver que o neutral é o que tem seu desempenho mais incoerente,isso pode ser decorrente de uma classificação errônea na base de dados original ou até palavras muito próximas nas outras classes fazendo com que o classificador tendesse para os resultados incorretos.

4.2. Aplicação das bases reais

Apos treinar o classificador, o passo final foi executar uma rotina em cada um atribuindo valores para cada um dos rótulos:

Negative -1
Neutral 0
Positive 1

Com esses rótulos foi possível montar um gráfico mostrando a media ,a cada três meses, dos sentimentos apontados por cada politico. De posse dessa tabela conseguimos chegar aos gráficos comparativos de emoções expressas por ambos durante suas campanhas que levaram a suas respectivas eleições. Os gráficos gerados por esse estudo encontram-se na seção de anexo.

5. Conclusão

Com base no que foi apresentado, o método de análise de sentimentos, por meio de aprendizagem de máquina, utilizando o algoritmo de naive bayes se mostrou utilizável quando buscamos essa análise de poucos rótulos e desconsiderando informações anteriores, coisa que foi necessário por se tratar de momentos distintos da historia de cada nação. Seu

percentual de acurácia mostrou-se aceitável, entretanto passível de melhora se estudado mais afundo as palavras que certos grupos sociais, nesse caso o grupo político utilizam, afinal, certas palavras que não foram vista pelo classificador, pois, a base de treinamento de ambos eram bases de treinamento de usuários da rede social twitter.

6. References

Todas as bases utilizadas nesse trabalho foram cedidas por usuarios do portal Kaggle e GitHUb, para acessar tal base basta acessar o repositorio no Github deste artigo: encurtador.com.br/mDHQY

7. Anexo

Table 1. 1º quarter

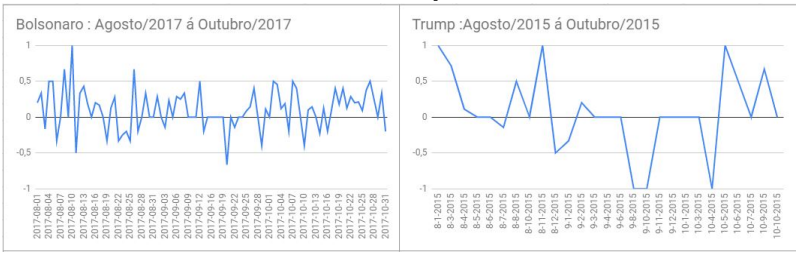

Table 2. 2º quarter

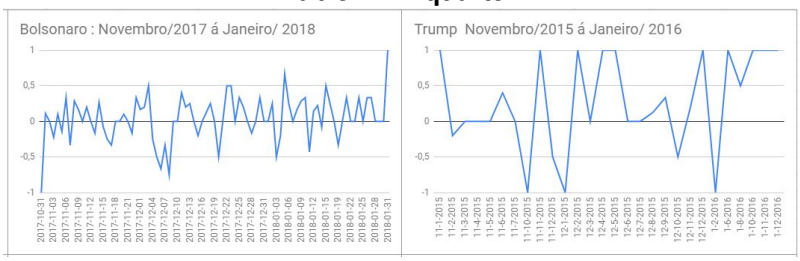

Table 3. 3º quarter

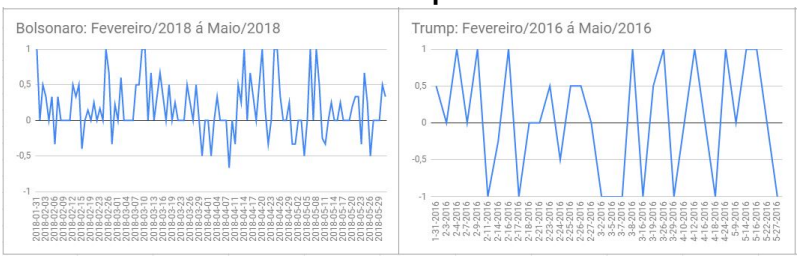

Table 4. 4º quarter