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

A área 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.

Foram utilizadas algumas ferramentas na construção dos projetos, foram elas,GoogleColab[1],Python[2],NLTK[3].

2.1. Base de Dados

Neste trabalho foram utilizadas quatro bases de dados distintas:

Table 1. Base de Dados Utilizadas

Base	Tipo	Link
1	Tweets em português	http://bit.do/tweetsPt
2	Tweets em inglês	http://bit.do/tweetsIn
3	Tweets do Jair Bolsonaro	http://bit.do/TweetsBolsonaro
4	Tweets do Donald Trump	http://bit.do/TweetsDonaldTrump

3. Métodos

3.1. Naive Bayes

Foi escolhido o Naive Bayes visto que é um algoritmo classificador probabilístico baseado no teorema estatístico de Bayes, sendo popular na área de aprendizagem de máquina, pois, caracteriza aquele 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.

[1] colab.research.google.com
[2] www.python.org
[3] www.nltk.org

3.2. Tratamento das Bases de Dados

Primeiro foi realizado uma limpeza nas colunas que compunham as base de dados, visto que algumas colunas possuíam dados irrelevantes, nas bases de treinamento e teste foram deixadas apenas as mensagens e os rótulos, já nas bases de dados das personalidades políticas, foram deixadas apenas as mensagem e as datas. Também foram removidos, quase que em sua totalidade, todos os emoticons, links e números. 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.

3.3. 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.

3.4. 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.

3.5. 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).

4.1. Testes de pré-processamento

Foram realizados testes de pré-processamento para saber escolher aquele que gerava o melhor resultado, os testes ocorreram de forma individual para cada classificador, assim sendo viável a escolha da melhor combinação por classificador. No classificador para Português foi escolhido o uso dos dois pré-processamento, o uso de Steaming e a retirada de Stopwords, pois essa combinação alcançou uma acurácia de 62,70%, enquanto no classificador para Inglês for escolhido a combinação que utilizou apenas da retirada dos Stopwords, visto que essa combinação alcançou uma acurácia de 68,24%.

Table 2. Testes de pré-processamento no classificador para Português

Combinação	Steaming	Stopwords	Acurácia
Combinação 1	X	X	62,70%
Combinação 2	X		47,44%
Combinação 3		X	61,43%

Table 3. Testes de pré-processamento no classificador para Inglês

Combinação	Steaming	Stopwords	Acurácia
Combinação 1	X	X	67,63%
Combinação 2	X		45,80%
Combinação 3		X	68,24%

Com isso após as modificações nos classificadores e seguindo pela melhor combinação e assim seguimos os experimentos com os seguintes valores.

Classificador em Inglês Obteve 62,70%
Classificador em Português Obteve 68,24%

4.2. 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.

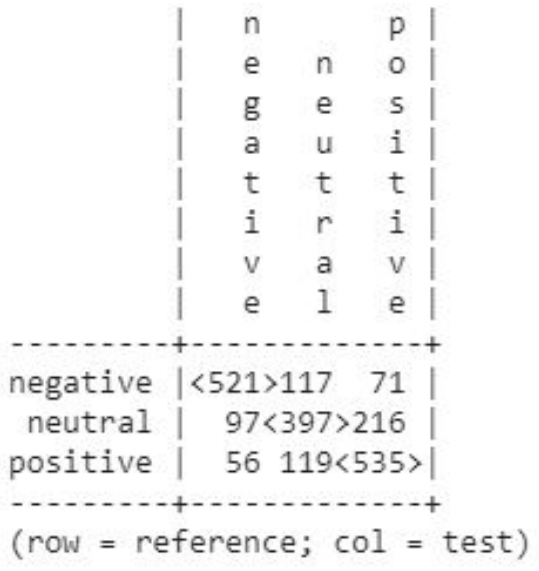

Figure 1. Matriz de Confusão do Classificador Inglês

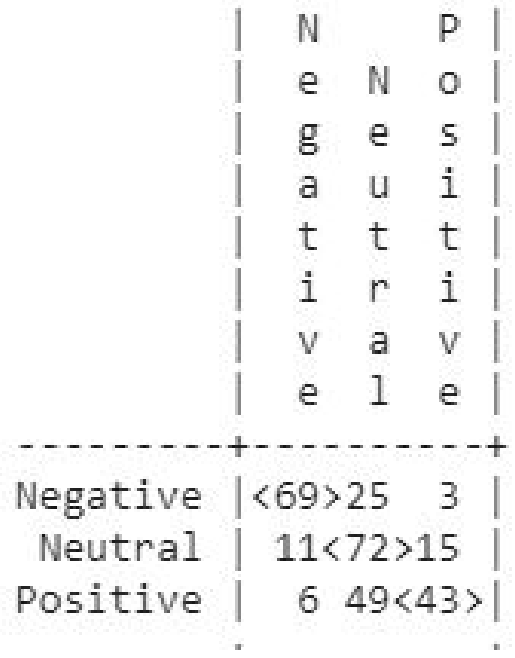

Figure 2. Matriz de Confusão do Classificador Português

Na figura 1 temos matriz de confusão do classificador em inglês que obteve uma alta taxa de acertos tanto no negativo quanto no positivo tendo apenas um alto nível de falso resultados quando avaliou palavras neutras. Já na matriz de confusão do classificador em português(figura 2) temos que a rotulação de frases positivos foi a de menor precisão acontecendo muitas rotulações equivocadas quanto a palavras neutras.

4.3. Aplicando nas bases Presidenciais

Após a realização do treino de cada classificador em sua melhor formação seguimos para aplicar a base dos políticos alvo do estudo para que possa ser feito a analise de suas postagens. Afim de conseguirmos quantificar em valores numérico as respostas , foi criado uma tabela de co-relação entre as classes utilizadas e alguns números .

Negative -1

Neutral 0

Positive 1

Feito isso foi a hora de montar uma tabela utilizando o Google Planilhas [4] e assim toda aqueles dados foram exportados para poder termos uma visão geral dos dados coletados.Na tabela a baixo teremos alguns resultados com a analise de dados resultante.

[4]docs.google.com/spreadsheets/

Table 4. Informações Extraídas

Informação	Bolsonaro	Trump
Sentimento predominante	Neutro	Neutro
Mês de mais postagens Positivas	Outubro/17	Agosto/15
Mês de mais postagens Negativas	Dezembro/17	Setembro/15
Mês de mais postagens Neutras	Outubro/17	Outubro/15
Porcentagem de postagens Positiva	23,26%	38,56
Porcentagem de postagens Negativa	12,87%	27,36%
Porcentagem de postagens Neutra	63,87%	34,08%

Logo após afim de comprar as variações durante as faixas temporais analisadas, criamos gráficos mostrando a variação dos sentimentos expressos no decorrer de cada faixa temporal, que foi de três meses dividindo assim o ano estudado em quatro etapas.

Table 5. 1º Trimestre

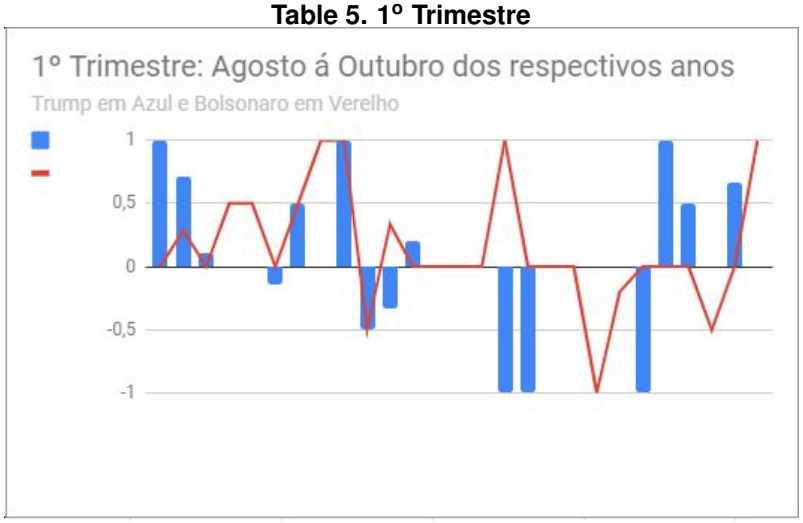

Table 6. 2º Trimestre

Table 7. 3º Trimestre

Table 8. 4ºTrimestre

Com base na união dos gráficos é possível visualmente ilustrar que por mais que tenhamos utilizado quantidade de dados diferentes em determinados meses a media que foi tirado das entradas demonstra-se ser a mesma na maioria dos pontos visualizados. Não é possível só com base no que foi coletado provar uma relação existente entre as mensagens dos indivíduos analisados por este trabalho,afinal tivemos bases maiores de um candidato em relação a outro , entretanto é capaz de mostrar um possível relação entre o que ambos queriam denotar durante esse ano antes e durante suas campanhas.

5. Trabalhos Futuros

Buscamos continuar afim de usar bases voltadas a cunho politico e quem sabe introduzir outros tipo de analise assim como a de sentido semântico das frases e seus caracteres especias e pontuações.

6. 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. A importância de ter realizado os testes de saber como melhor realizar a limpeza dos dados mostrou que é possível realizar duas formas de pré-processamento de dados afim de atingir um melhor potencial de cada classificador.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.

7. Referências

Todas as bases utilizadas nesse trabalho foram cedidas por usuários do portal Kaggle e GitHUb, tais bases encontram-se neste link: encurtador.com.br/mDHQY