# Categorização de desvios de conduta a partir dos dados de *tweets* do presidente Jair Bolsonaro.

***Abstract.*** *This article has as main objective to analyze all the tweets until July 2020, of the current president of Brazil, Jair Messias Bolsonaro and, to relate the behavior of the texts in relation to the pre, during and post candidacy period. For that, he used the Naive Bayes, SVM and Decision Tree algorithms to make the predictions.*

***Resumo.*** *Este artigo tem como principal objetivo analisar todos os tweets até o Julho de 2020, do atual presidente do Brasil, Jair Messias Bolsonaro e, relacionar o comportamento dos textos em relação ao período pré, durante e pós candidatura. Para isso utilizou dos algoritmos Naive Bayes, SVM e Árvore de decisão para fazer as predições.*

## 1. Introdução

O presente trabalho tem como objetivo analisar o *dataset* referente aos *tweets* do atual presidente da República, Jair Messias Bolsonaro. O conjunto de dados, pertencente ao período de março de 2010 à julho de 2020, possui 9351 entradas e 5 atributos, na Tabela 1 é possível verificar as colunas existentes e seus respectivos tipos. As tabelas de 2 à 6 detalham as informações pertinentes a cada atributo.

Essa base de dados foi escolhida pela amplitude de meios para aplicar as técnicas propostas de reconhecimento de padrões, podendo ser avaliado em diferentes contextos. Além do acesso facilitado ao conjunto de dados, o período escolhido condiz com os anos em que o atual presidente obteve parcela considerável dos holofotes, tanto na mídia, no contexto político, quanto na rede social propriamente dita, o Twitter.

A separação datada tem como propósito analisar as informações em três épocas: antes da eleição, durante o ano eleitoral e após. Como não há conhecimento prévio acerca dos resultados, a análise de cada técnica será utilizada no processo de entendimento de possíveis padrões de comportamento, sendo o fruto deste trabalho estritamente um relatório técnico. A variação de tais padrões de comportamento poderá ilustrar a coesão e coerência dos de determinados posicionamentos do político em questão no início e no final deste período.

Procedimentos distintos serão aplicados, objetivando separar as informações a partir de dados que caracterizem as mensagens postadas pelo mesmo como posturas incompatíveis ou inaceitáveis com o cargo que ocupa, através de conteúdo ofensivo, discriminatório e antidemocrático, fazendo uma análise no que pode acarretar tal comportamento para a sociedade, além da data tornar possível entender em qual contexto esses posicionamentos estavam inseridos. Tal análise poderá ser utilizada como referência para trabalhos futuros para traçar possíveis paralelos com outros políticos que também possam apresentar posturas questionáveis nos cargos que ocupam.

Tabela 1 - Descrição das colunas da base de dados

| Coluna | Tipo |
|---|---|
| Data | *DateTime* |
| Texto do *tweet* | *String* |
| Número de curtidas | *Integer* |
| Número de *retweets* | *Integer* |
| Link para o *tweet* | *String* |

Fonte: https://www.kaggle.com/lgmoneda/jair-bolsonaro-twitter-data

Tabela 2 - Detalhes da coluna: Dia em que foi tweetado

| Válidos | Incompatíveis | Ausentes | Mínimo | Média | Máximo |
|---|---|---|---|---|---|
| 9351 | 0 | 0 | 31/03/2010 | 07/06/2018 | 31/07/2020 |

Fonte: https://www.kaggle.com/lgmoneda/jair-bolsonaro-twitter-data

Tabela 3 - Detalhes da coluna: Texto do *tweet*

| Válidos | Incompatíveis | Ausentes | Únicos | Mais comuns |
|---|---|---|---|---|
| 9351 | 0 | 0 | 9334 | Brasil |

Fonte: https://www.kaggle.com/lgmoneda/jair-bolsonaro-twitter-data

Tabela 4 - Detalhes da coluna: Número de curtidas

| Válidos | Incompatíveis | Ausentes | Média | Desvio Padrão | Total |
|---|---|---|---|---|---|
| 9351 | 0 | 0 | 16200 | 22500 | 224000 |

Fonte: https://www.kaggle.com/lgmoneda/jair-bolsonaro-twitter-data

Tabela 5 - Detalhes da coluna: Número de *retweets*

| Válidos | Incompatíveis | Ausentes | Média | Desvio Padrão | Total |
|---|---|---|---|---|---|
| 9351 | 0 | 0 | 3060 | 4080 | 91100 |

Fonte: https://www.kaggle.com/lgmoneda/jair-bolsonaro-twitter-data

Tabela 6 - Detalhes da coluna: Link para o *tweet*

| Válidos | Incompatíveis | Ausentes | Únicos |
|---|---|---|---|
| 9351 | 0 | 0 | 9351 |

Fonte: https://www.kaggle.com/lgmoneda/jair-bolsonaro-twitter-data

## 2. Metodologia

A pesquisa, segundo Gerhardt e Silveira (2009) “possibilita uma aproximação e um entendimento da realidade a investigar. A pesquisa é um processo permanentemente inacabado. Processa-se por meio de aproximações sucessivas da realidade, fornecendo-nos subsídios para uma intervenção no real”.

A modalidade de pesquisa escolhida para esse trabalho tem uma abordagem inicialmente quantitativa, com a divisão dos dados em três períodos, depois qualitativa, classificando os *tweets*. Quanto a sua natureza, é uma pesquisa básica, sem aplicação

prática envolvida. Quanto aos seus objetivos, é uma pesquisa exploratória com o objetivo de proporcionar maior familiaridade com o problema, a fim de construir hipóteses. Por fim, quanto aos seus procedimentos, através de uma pesquisa documental sobre os *tweets*, será feito uma pesquisa ex-post-facto, investigando possíveis relações de causa e efeito entre atitudes da população e o atual presidente.

## 2.1. Coleta de Dados e pré-processamento

O download da base de dados foi feito pelo site Kaggle, no dia 22 de agosto de 2020, com *tweets* até agosto de 2020 e título Dados do Twitter de Jair Bolsonaro: *Tweets* de Jair Bolsonaro, Presidente da República Federativa do Brasil.

Inicialmente, o *dataset* foi dividido em três arquivos que serão analisados individualmente pelos algoritmos escolhidos, baseado pela data. A partir disso, de 2010 à 2017, foi realocado no arquivo referente ao período de pré ano eleitoral, já o ano de 2018 refere-se ao ano eleitoral e, por fim, de 2019 à 2020, o arquivo de pós eleição. Cada arquivo contém, respectivamente: 3144, 1984 e 4223; para igualar a apuração de acertos do algoritmo, foi delimitado a análise e classificação manual em torno de 500 entradas aleatórias de cada arquivo, com o objetivo de analisar a taxa de acertos do algoritmo.

Para a classificação da base de dados, foi utilizada apenas a coluna Texto do *tweet* e, uma coluna Sentimento foi adicionada ao novo *dataset*, com os valores: ofensivo, neutro, positivo e antidemocrático. Essa coluna foi classificada, inicialmente, como resultado da análise de todos integrantes do grupo.

Na categorização de textos ofensivos foram considerados termos discriminatórios e de cunho preconceituoso, assim como conteúdos negativos em um contexto geral. Para os antidemocráticos, *fake news* e ataques a outros governos, assim como o próprio sistema político nacional, além de ataques a liberdade de expressão e distribuição de notícias midiáticas foram separados nesta lista.

Ao analisar a categoria neutro, incluiu-se os que continham textos que não alcançaram outras categorias, com conteúdo irrelevante de análise. Por fim, os categorizados como positivos foram obtidos através de conteúdos postados que apresentavam medidas de melhoramento para a gestão ou grupos específicos e, que, principalmente, foram feitos sem termos ou ataques ofensivos a outros grupos, não englobando outras categorias de classificação.

Tabela 7 - Separação e classificação de *tweets*

| Data | Texto | Sentimento |
|---|---|---|
| 2013-03-07 | RT @NubladoVentania: Após ser agredido na CDDH por gayzistas,Bolsonaro diz: "Voltem p o zoológico...bando d viadada". http://bit.ly/YEx7Is | ofensivo |
| 2018-08-14 | Mais uma vez número um nos trending topics no twitter com a hashtag #BolsonaroNaRecord . Obrigado pelo apoio 👍. Boa noite a todos! https://t.co/p1eT4N7QQn | neutro |

| | | |
|---|---|---|
| 2020-07-25 | "A- Agora às 18hs, juntamente com a @AdvocaciaGeral, entrei com uma ADIn no STF visando ao cumprimento de dispositivos constitucionais.<br>- Uma ação baseada na clareza do Art. 5°, dos direitos e garantias fundamentais. (Continua na próxima postagem)." | positivo |
| 2016-03-18 | MAIS UM COMUNISTA DO PT APAVORADO CONOSCO! http://tinyurl.com/gtg655w | antidemocrático |

Fonte: Os autores, 2020

## 2.2. Classificação por algoritmos

De acordo com o cenário proposto, serão implementados três algoritmos, sendo eles o Máquina de Vetores de Suporte (SVM), Árvore de Decisão e Naive Bayes, todos utilizando da linguagem de programação Python.

Uma máquina de vetores de suporte (SVM) é um tipo de algoritmo de classificação de aprendizado de máquina supervisionado. Existe desde o SVM simples, até o mais avançado, conhecido como Kernel SVM. Nesse foi utilizado o SVM simples, ou também chamado de linear, no qual foi codificado usando a biblioteca SVM com Scikit-Learn, além do auxílio das bibliotecas pandas e numpy (OLIVEIRA; PRUDÊNCIO, 2010).

Para Campos (2017), Árvores de decisão são métodos de aprendizado de máquinas supervisionado, muito utilizados em tarefas de classificação e regressão. Árvores, de um modo geral em computação, são estruturas de dados formadas por um conjunto de elementos que armazenam informações chamadas nós.

Além disso, toda árvore possui um nó chamado raiz, que possui o maior nível hierárquico (o ponto de partida) e ligações para outros elementos, denominados filhos. Esses filhos podem possuir seus próprios filhos que por sua vez também possuem os seus. O nó que não possui filho é conhecido como nó folha ou terminal.

Segundo França e Oliveira (2013) “Naive Bayes é um método de aprendizagem probabilística baseado no teorema de Bayes. Este modelo assume que a probabilidade de presença ou ausência de uma característica particular de uma classe [...]”.

Os algoritmos trabalham com o mesmo pré-processamento, primeiro é necessário escolher o intervalo de tempo desejável para a análise (pré, durante ou pós ano eleitoral), depois a função *pre_processamento* faz a manipulação dos dois arquivos, um direcionado ao treinamento e o outro para os testes. A pasta do projeto possui no total seis arquivos, dois de cada período.

Para ler os arquivos, utilizou-se a biblioteca pandas (*pd.read_csv*) e chamou a função *head* para retornar as primeiras entradas. Após, exclui-se os *stop words* dos textos, em português, palavras de parada, são palavras que podem ser consideradas irrelevantes para o conjunto de resultados, através da função de NLTK.

O arquivo de treinamento possui apenas os *tweets* com sentimento já classificado e, consequentemente, o arquivo de teste ainda será preenchido pelo algoritmo. A função *loc* é utilizada dentro da variável de teste para percorrer os valores da coluna sentimento e atribuir valores de 0 à 3 para cada tipo de sentimento identificado do arquivo. Em seguido, ocorre a chamada da função de *train_test_split*, definindo as colunas de treinamento.

Cada algoritmo possui uma função específica, para o funcionamento do Naive Bayes utiliza-se o *CountVectorizer* que associa-se através da função Pipeline com o *MultinomialNB()*, utilizado para contagens do tipo. A função para o SVM é denominada *svm* e para a Árvore de decisão, *tree*. Todas acompanhadas de suas funcionalidades.

O processo final é igual para todos os algoritmos, a fim de comparar os resultados, sendo assim a função *fit* é aplicada com os recursos de treinamento na variável que recebeu as ações anteriores e, por fim, esta variável chama a função score nos arquivos de teste que delimitaram sua taxa de acertos.

Além disso, a função *predict* é o responsável pela geração dos arquivos finais, por ele são mostrados os possíveis valores para cada linha de texto do arquivo de teste, previsto pelos algoritmos. Por essa função é possível fazer a contagem do resultado.

## 3. Resultados e discussão

Após a compilação dos algoritmos, é importante ressaltar que os resultados não são valores estáticos. Observe na Figura 1 que é possível escolher o período desejado para fazer a análise e, é mostrado o total de dados do arquivo de treino da opção escolhida, no caso foi o intervalo antes da candidatura, com 507 entradas de treino.

Figura 1 - Quantidade de dados treinados no arquivo de pré ano eleitoral

```
--------------------------------------------------------------------------
Qual período dos tweets de Jair Bolsonaro você deseja analisar?
 [1] Pré candidatura
 [2] Durante candidatura
 [3] Pós candidatura
Digite apenas uma opção: 1
--------------------------------------------------------------------------
Para esse treinamento foram utilizados 507 tweets:
antidemocrático    249
neutro             160
ofensivo            89
positivo             9
Name: sentimento, dtype: int64
```

Fonte: Os autores, 2020.

Os resultados mostram as taxas de acurácia de cada algoritmo e a quantidade de *tweets* que foram previstos para cada sentimento, como na Figura 2.

Figura 2 - Resultados pré ano eleitoral

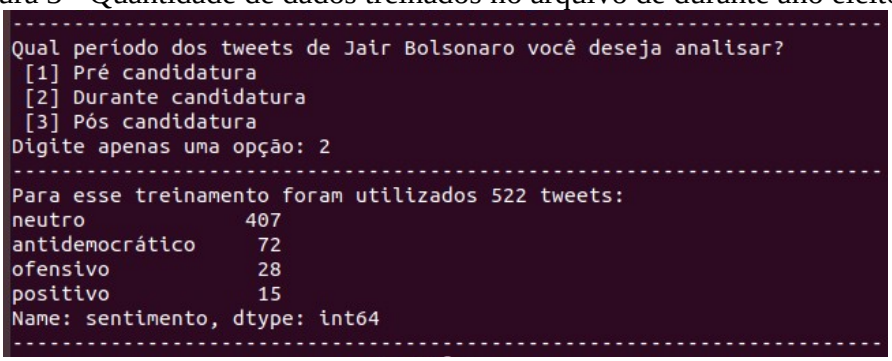
```
--------------------------------------------------------------------------
                            PREDIÇÃO
--------------------------------------------------------------------------
Naive Bayes

Esse algoritmo preveu, dentre 2637 tweets, a quantidade a seguir:
antidemocrático    1654
neutro              871
ofensivo            112
Name: sentimento, dtype: int64

Acurácia:  0.7086614173228346
--------------------------------------------------------------------------
SVM

Esse algoritmo preveu, dentre 2637 tweets, a quantidade a seguir:
antidemocrático    1898
neutro              700
ofensivo             39
Name: sentimento, dtype: int64

Acurácia:  0.6850393700787402
--------------------------------------------------------------------------
Árvore de decisão

Esse algoritmo preveu, dentre 2637 tweets, a quantidade a seguir:
antidemocrático    2637
Name: sentimento, dtype: int64

Acurácia:  0.5039370078740157
```

Fonte: Os autores, 2020.

A Figura 2 apresenta as respostas para os arquivos do período pré eleitoral, note que Naive Bayes obteve uma taxa de acertos em torno de 70,87%, já o SVM obteve 68,50% de eficiência, enquanto a Árvore de Decisão obteve 50,39%, a menos eficaz entre os três algoritmos, é importante ressaltar que o período em questão possuiu 2637 número de entradas para classificação.

Para o período equivalente ao ano eleitoral, em 2018, a quantidade de *tweets* utilizados para treino foram 522, como mostra a Figura 3.

Figura 3 - Quantidade de dados treinados no arquivo de durante ano eleitoral




Fonte: Os autores, 2020.

A taxa de acertos na data mencionada foi em torno de 83,96% para o Naive Bayes, o SVM teve eficiência de aproximadamente 85,50% e a Árvore de Decisão obteve uma melhora já que neste conseguiu alcançar 83,21%, sendo o Naive Bayes o menos eficaz entre todos, este conjunto de dados também contém o menor montante analisado para os testes.

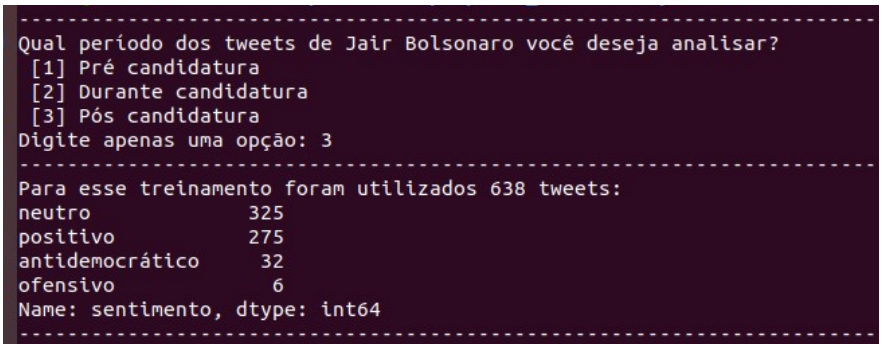
Figura 4 - Resultados durante ano eleitoral


Fonte: Os autores, 2020.

Por fim, para o período de pós eleição, de 2019 até 2020, a quantidade de *tweets*

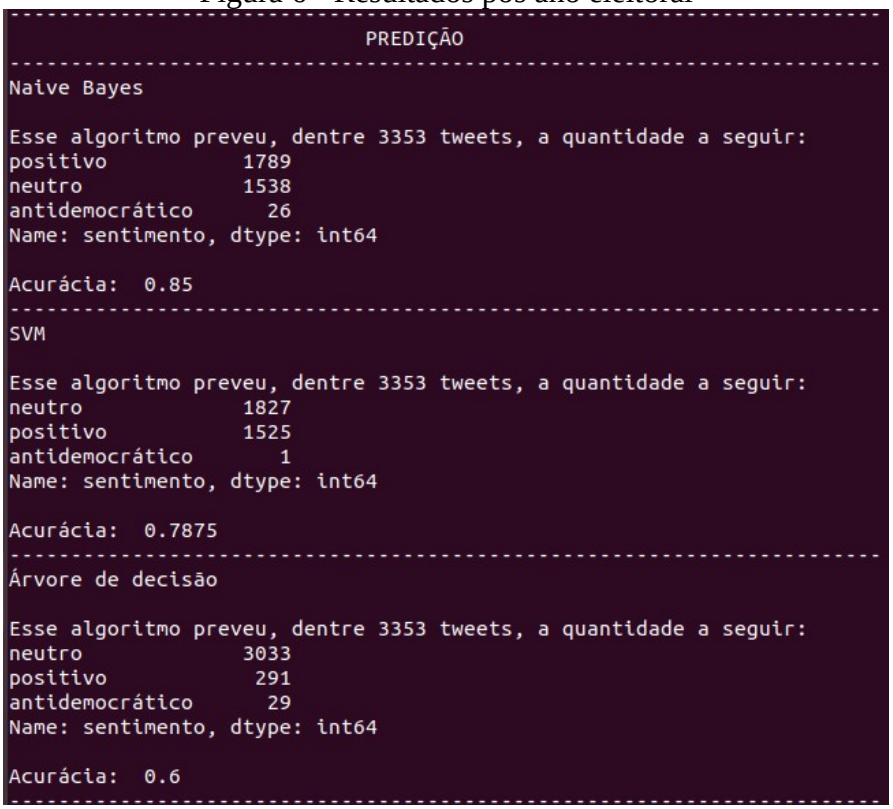
analisados fora 638 (Figura 5) e, a taxa foi em torno de 85% para o Naive Bayes, para o SVM 78,75% e, o menos eficaz do grupo, a Árvore de decisão com 60%. (Figura 6)

Figura 5 - Quantidade de dados treinados no arquivo de durante ano eleitoral




Fonte: Os autores, 2020.

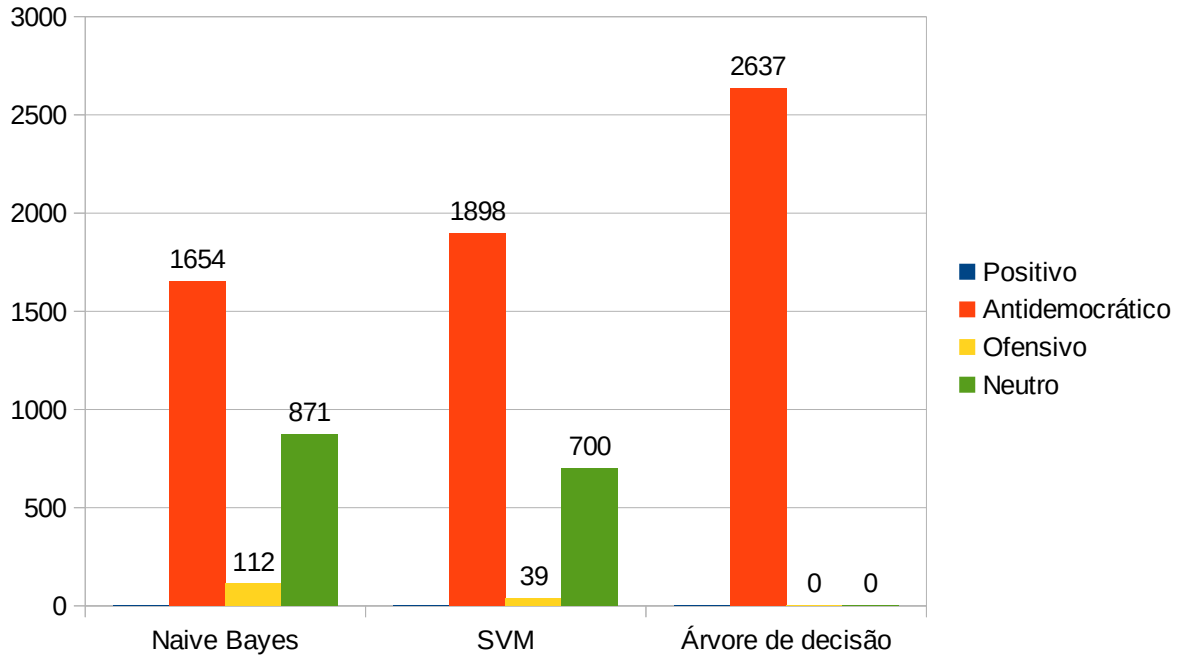
Figura 6 - Resultados pós ano eleitoral




Fonte: Os autores, 2020.

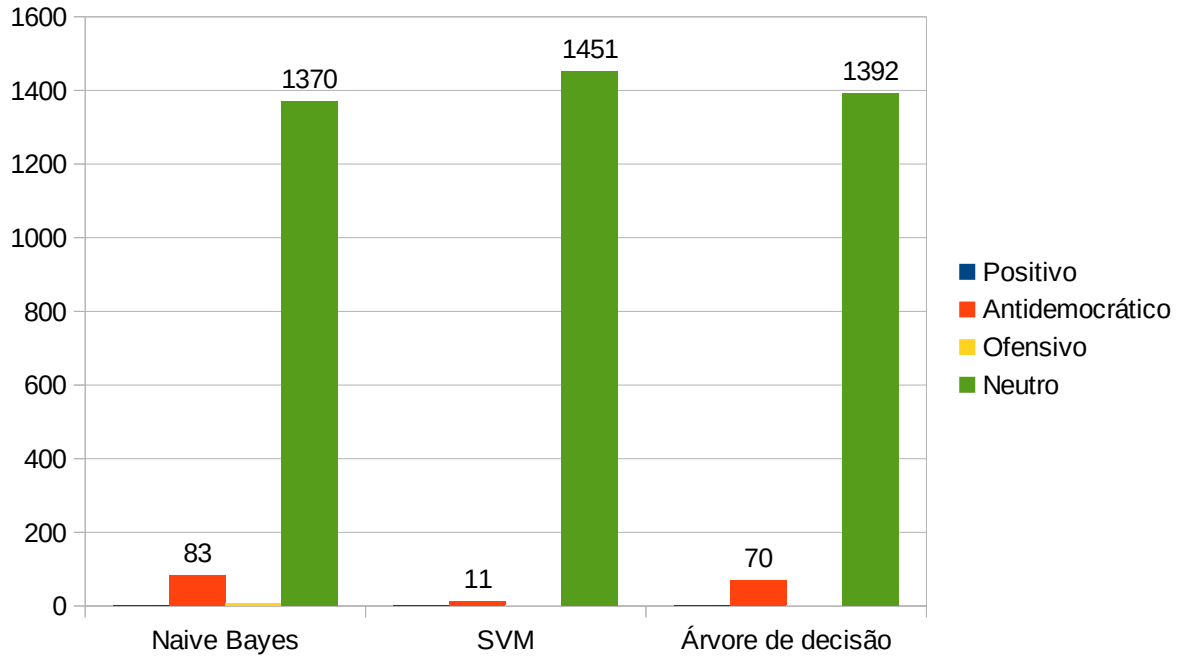
Cada algoritmo gerou três arquivos referentes a predição dos testes. De acordo com as predições, notou-se que no período pré ano eleitoral, que a maioria dos textos se destacaram pelo sentimento antidemocrático, como expõe a Figura 7. A Figura 8 exibi as predições durante o período da candidatura, nessa predominou as mensagens consideradas neutras pelos algoritmos. E por fim, a fase atual que é pós o ano eleitoral, nessa teve um aumento significativo nos comentários positivos, observe a Figura 9.

Figura 8 – Resultado dos dados de teste no arquivo pré ano eleitoral



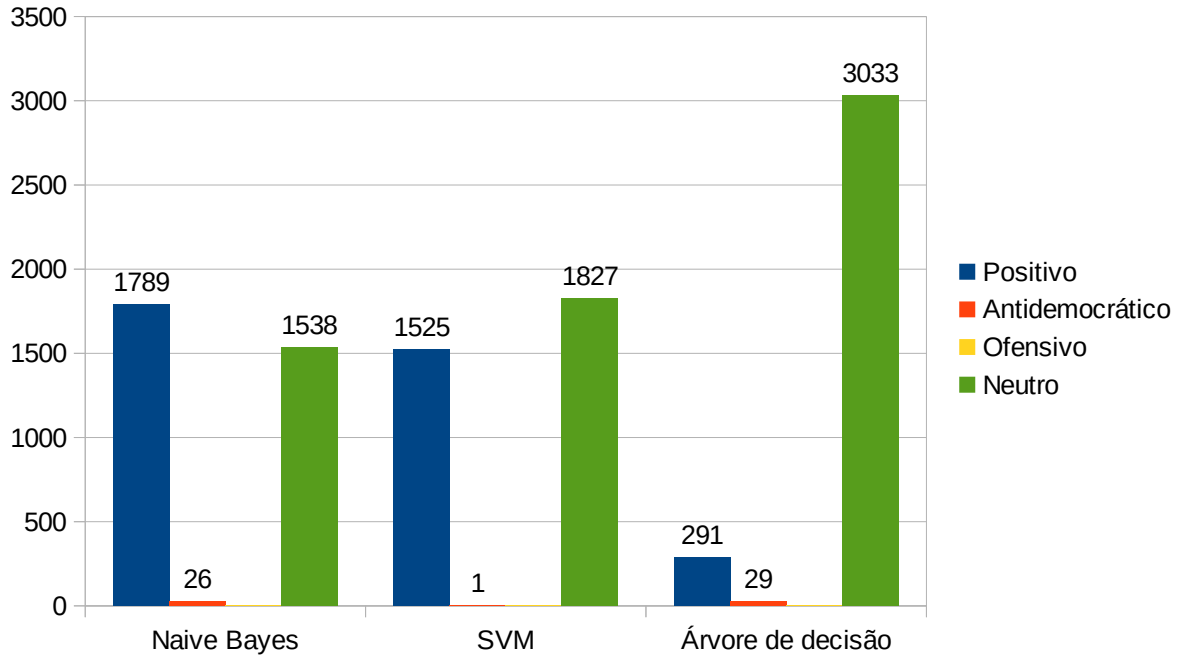

Fonte: Os autores, 2020.

Figura 8 - Resultado dos dados de teste no arquivo durante ano eleitoral




Fonte: Os autores, 2020.

Figura 9 - Resultado dos dados de teste no arquivo pós ano eleitoral




Fonte: Os autores, 2020.

## 4. Conclusão

O objetivo de analisar os *tweets* do presidente Jair Bolsonaro, em três diferentes períodos, foi atingido. Utilizou-se três algoritmos e notou-se que a Árvore de Decisão se mostrou menos eficiente para o problema, enquanto o Naive Bayes foi mais eficiente dentre eles, tendo o SVM como intermediário. Além disso, segundo as predições, os *tweets* pré ano eleitoral era predominantemente antidemocrático, enquanto no período eleitoral ressalva *tweets* de cunho neutro, e atualmente os *tweets* são mais positivos.

## Referências

CAMPOS; R.; Árvores de Decisão. 2017. Disponível em:https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 09 de set. 2020.

FRANÇA, T. C.; OLIVEIRA, J. Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. 2013. Disponível em: http://www.each.usp.br/digiampietri/BraSNAM/2014/p11.pdf. Acesso em: 09 de set. 2020.

GERHARDT, T. E.; SILVEIRA, D. T. Métodos de pesquisa. Porto Alegre: Editora da UFRGS, 2009.

OLIVEIRA, G. M. O.; PRUDENCIO, R. B. C. Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado. 2010. Disponível em: https://www.cin.ufpe.br/~tg/2010-2/gmoj.pdf. Acesso em: 09 de set. 2020.