Avaliação de métodos de mineração de textos aplicados à detecção de Fake News eleitorais brasileiras

Evaluation of Text Mining Methods Applied to The Detection of Brazilian Electoral Fake News

Evaluación de Métodos de Minería de Texto Aplicados a la Detección de Noticias Falsas Electorales Brasileñas

Caio Vinícius Meneses Silva
Mestre em Ciência da Computação
Universidade Federal de Sergipe
caio_vms@outlook.com

Methanias Colaço Júnior
Pós-doutorado em Gestão e Doutor em Computação
Universidade Federal de Sergipe
mjrse@hotmail.com

Raphael Silva Fontes
Mestre em Ciência da Computação
Universidade Federal de Sergipe
raphaelf.ti@gmail.com

Resumo

A evolução dos meios de comunicação tem contribuído com a disseminação de *fake news*, principalmente após o surgimento das redes sociais digitais. A velocidade com que estas notícias se espalham tornaram inviável a checagem manual desse imenso volume de dados. Diante deste contexto, trabalhos em diversas áreas têm sido realizados a fim de tentar minimizar os danos causados pela proliferação das denominadas *fake news*. O objetivo deste trabalho é avaliar a eficácia dos métodos mais utilizados para verificar correspondência de textos, no contexto da detecção de *fake news*, tendo como base as eleições presidenciais brasileiras de 2018, bem como fazendo um comparativo com os resultados da eleição norte-americana de 2016, publicados na literatura. Adicionalmente, uma visão geral das *fake news* por seguidores de cada candidato é apresentada. Foi planejado e executado um experimento controlado, para

comparar a eficácia dos métodos selecionados. Os métodos TF-IDF e BM25 se destacaram nesse contexto, possuindo, estatisticamente e respectivamente, médias similares de Acurácia (79,86% e 79,00%), Precisão (79,97% e 78,76%), Sensibilidade (78,97% e 76,05%) e Medida-F1 (79,47% e 77,38%). A eficácia foi similar à do contexto norte-americano, no qual o BM25 alcançou uma Acurácia de 79,99%. Além disso, considerando o universo de notícias checadas disponível, o período analisado e uma margem de erro de 3,5%, evidenciou-se que houve divulgação de *fake news* por ambos os lados e que seguidores do candidato Jair Bolsonaro (PSL) foram responsáveis por 62,25% dos *tweets* relacionados a *fake news*, contra 37,75% dos seguidores do candidato Fernando Haddad (PT). No que diz respeito às contas excluídas da rede social em um curto espaço de tempo, 59,96% eram de seguidores do candidato do PSL e 40,04% de seguidores do candidato do PT. A divulgação de *fake news* nem sempre implica intenção, podendo implicar apenas um engajamento maior por parte de alguns seguidores.

Palavras-chave: Eleições. Experimentação. *Fake news.*

Abstract

The evolution of the media has contributed to the spread of false news, especially after the emergence of digital social networks. The speed with which this news spread made it impossible to manually check this huge amount of data. In this context, work in several areas has been carried out in order to try to minimize the damage caused by the proliferation of so-called fake news. The objective of this work is to evaluate the effectiveness of the most used methods to check correspondence of texts, in the context of detecting false news, based on the Brazilian presidential elections of 2018, as well as making a comparison with the results of the US election. 2016, published in the literature. Additionally, an overview of the *fake news* by followers of each candidate is presented. A controlled experiment was planned and executed to compare the effectiveness of the selected methods. The TF-IDF and BM25 methods stood out in this context, having, statistically and respectively, similar averages of Accuracy (79,86% and 79,00%), Precision (79,97% and 78,76%), Sensitivity (78,97% and 76,05%) and Measure-F1 (79,47% and 77,38%). The effectiveness was similar to that of the North American context, in which the BM25 achieved an Accuracy of 79,99%. Furthermore, considering the universe of checked news available, the analyzed period and a margin of error of 3,5%, it was evident that *fake news* were disclosed by both sides and that followers of the candidate Jair Bolsonaro (PSL) were responsible for 62,25% of tweets related to fake news, against 37,75% of followers of candidate Fernando Haddad (PT). With regard to accounts deleted from the social network in a short time, 59,96% were followers of the PSL candidate and 40,04% of followers of the PT candidate. The dissemination of fake news does not always imply intention, and may only imply greater engagement by some followers.

Keywords: Elections. Experimentation. *Fake news*.

Resumen

La evolución de los medios ha contribuido a la difusión de noticias falsas, especialmente tras la aparición de las redes sociales digitales. La velocidad con la que se difundió esta noticia hizo imposible verificar manualmente esta enorme cantidad de datos. En este contexto, se ha trabajado en varios ámbitos para tratar de minimizar el daño causado por la proliferación de las llamadas *fake news*. El objetivo de este trabajo es evaluar la efectividad de los métodos más utilizados para verificar la correspondencia de los textos, en el contexto de la detección de noticias falsas, con base en las elecciones presidenciales brasileñas de 2018, así como realizar

una comparación con los resultados de las elecciones estadounidenses. 2016, publicado en la literatura. Además, se presenta un resumen de las *fake news* por seguidores de cada candidato. Se planificó y ejecutó un experimento controlado para comparar la efectividad de los métodos seleccionados. Los métodos TF-IDF y BM25 se destacaron en este contexto, teniendo, estadísticamente y respectivamente, promedios similares de Exactitud (79,86% y 79,00%), Precisión (79,97% y 78,76%), Sensibilidad (78,97% y 76,05%) y Medida-F1 (79,47% y 77,38%). La efectividad fue similar a la del contexto norteamericano, en el que el BM25 logró una Precisión del 79,99%. Además, considerando el universo de noticias comprobadas disponibles, el período analizado y un margen de error del 3,5%, se evidenció que las *fake news* fueron divulgadas por ambas partes y que los seguidores del candidato Jair Bolsonaro (PSL) eran responsables de El 62,25% de los tuits relacionados con *fake news*, frente al 37,75% de los seguidores del candidato Fernando Haddad (PT). En cuanto a las cuentas eliminadas de la red social en poco tiempo, el 59,96% eran seguidores del candidato del PSL y el 40,04% de seguidores del candidato del PT. La difusión de noticias falsas no siempre implica intención, y solo puede implicar un mayor compromiso por parte de algunos seguidores.

Palabras clave: Extracción de textos. Experimentación. *Fake news*.

1. Introdução

Desde a popularização dos *smartphones*, o número de pessoas que utilizam redes sociais digitais tem aumentado a cada dia. Juntas, plataformas tais como *Facebook, WhatsApp* e *Twitter* possuem cerca de 4 bilhões de clientes ao redor do mundo (STATISTA, 2019). Este fenômeno alterou o modo como notícias são publicadas e consumidas, portanto, checar notificações, enviar e receber conteúdo por meio destas plataformas tornou-se uma tarefa rotineira. Todas estas interações realizadas por essa enorme quantidade de usuários de todo o mundo geram uma imensa massa de dados, comumente chamada de *Big Data* (CONROY; VICTORIA; YIMIN, 2015).

Como consequência dessa nova maneira de acesso à informação e desse aumento do volume de dados, o alcance a essas informações também se expandiu. Se, por um lado, o acesso quase imediato ao que acontece no mundo é algo extremamente útil, em contrapartida, a disseminação de conteúdo falso se apresenta como uma praga digital, pois, diante da velocidade com que se propagam, checar a veracidade de notícias tem se tornado uma tarefa extremamente complexa e humanamente quase inviável (CIAMPAGLIA; PRASHANT *et al.*, 2015).

Enxergando o potencial que estes novos meios de comunicação possuem para transmitir informações, métodos de análise de dados e testes de personalidade baseados em atividades de redes sociais têm sido utilizados para produzir e direcionar *fake news* a fatias altamente específicas da população, muitas vezes, visando gerar influência nos mais diversos segmentos da sociedade, a exemplo da política (ALLCOTT; GENTZKOW, 2017).

No entanto, a disseminação de conteúdo falso não é um fenômeno inédito, tampouco recente na história da humanidade. Existem relatos de alguns acontecimentos ao longo da história, como, por exemplo, o uso de propaganda por jornalistas na Primeira Guerra Mundial, que culminaram em novas normas de objetividade e equilíbrio jornalístico (DAVID; MATTHEW *et al.*, 2018). Nas mídias sociais digitais, tal fenômeno, agora chamado de *fake news*, encontrou um novo ambiente propício para se espalhar em escalas mundiais, causando sérios prejuízos à sociedade.

Desde 2016, a menção ao termo *fake news* aumentou em 365%, tornando-o a palavra do ano de 2017 (COLLINS, 2017). Traduzido do inglês, *fake news* significa "notícia falsa", todavia, o que caracteriza este termo com mais precisão, além de serem notícias propositalmente falsas, são as intenções obscuras existentes na divulgação massiva destas histórias falsas na era da internet, comumente usadas como forma de manipular as massas e suas opiniões públicas em encontro de um interesse específico.

Nos últimos anos, eventos políticos têm sido pautados por uma guerra virtual, cujo palco são as redes sociais, a exemplo das eleições presidenciais dos EUA em 2016 e do Brasil em 2018 (RUEDIGER; GRASSI *et al.*, 2017). Durante o período eleitoral, esse ambiente tem se tornado um campo de batalha altamente estratégico, no qual candidatos e apoiadores são ativamente envolvidos em fazer campanha, expressar suas opiniões e divulgar conteúdo, muitas das vezes falsos.

Para tentar mitigar os danos causados nos mais diversos seguimentos da sociedade, as redes sociais, que são os principais meios de propagação de *fake news,* têm tomado algumas medidas. O *Facebook*, por exemplo, criou um mecanismo com o qual é possível sinalizar uma publicação como falsa. Desta forma, o alcance da publicação é reduzido e o autor recebe uma advertência (ROCHLIN, 2017). O *WhatsApp*, hoje pertencente ao *Facebook*, decidiu estabelecer um limite para mensagens encaminhadas com muita frequência. Antes, o cliente poderia compartilhá-la com até cinco conversas de uma única vez, desde abril de 2020, a mensagem só poderá ser encaminhada para uma conversa por vez (WHATSAPP, 2020). O *Twitter* também anunciou, em 2018, um conjunto de regras mais rígidas para conter o avanço das *fake news* (TWITTER, 2019). Devido ao seu potencial de circulação de conteúdo jornalístico, o micro blog tem sido usado como parte estratégica de divulgação de informações falsas.

Fora do ambiente das redes sociais, outras ferramentas, tais como as agências de checagem de fatos, ou *fact-checking*, também têm auxiliado no combate às *fake news*. O *fact-checking* confronta histórias com dados, pesquisas e registros e é também uma forma de qualificar o debate público, por meio da apuração jornalística, além de averiguar o grau de veracidade das informações (SPINELLI; ALMEIDA SANTOS, 2018).

Todos esses esforços visam mitigar as graves consequências que as *fake news* podem e têm causado à sociedade, fomentando diversas linhas de pesquisa que mesclam os esforços manuais de jornalistas compromissados com a verdade e técnicas de Inteligência Artificial, e que estão consciente da necessidade de ferramentas que venham contribuir para a determinação da autenticidade dessas informações de uma forma cada vez mais automática.

Neste contexto, por meio da combinação do conhecimento gerado por agências de checagem de fatos com técnicas automáticas e inteligentes de análise de dados, o objetivo deste trabalho foi realizar um experimento para avaliar a eficácia dos métodos mais utilizados para verificar correspondência de textos, no contexto da detecção de *fake news*, tendo como base as eleições presidenciais brasileiras de 2018, bem como fazendo um comparativo com os resultados da eleição norte-americana de 2016, publicados na literatura. Adicionalmente, uma visão geral das *fake news* por seguidores de cada candidato foi apresentada. Os resultados evidenciaram que os métodos TF-IDF e BM25 se destacaram nesse contexto, possuindo, estatisticamente e respectivamente, médias similares de Acurácia (79,86% e 79,00%), Precisão (79,97% e 78,76%), Sensibilidade (78,97% e 76,05%) e Medida-F1 (79,47% e 77,38%).

Desta forma, a eficácia foi similar à do contexto norte-americano, no qual o BM25 alcançou uma Acurácia de 79,99%. Além disso, considerando o universo de notícias checadas disponível, o período analisado e uma margem de erro de 3,5%, evidenciou-se que houve divulgação de *fake news* por ambos os lados e que seguidores do candidato Jair Bolsonaro, do Partido Social Liberal (PSL), foram responsáveis por 62,25% dos tweets relacionados a *fake news*, contra 37,75% dos seguidores do candidato Fernando Haddad, do Partido dos Trabalhadores (PT). No que diz respeito às contas excluídas da rede social em um curto espaço de tempo, 59,96% eram de seguidores do candidato do PSL e 40,04% de seguidores do candidato do PT.

Para uma melhor compreensão de como obtivemos os resultados do experimento, o trabalho foi estruturado da seguinte forma. A Seção 2 traz uma base teórica dos métodos utilizados por este estudo. A Seção 3 é dedicada aos trabalhos relacionados. Na Seção 4,

descreve-se a metodologia utilizada na condução do trabalho. A Seção 5 aborda a definição e o planejamento do experimento, e, na Seção 6, relata-se a operação deste. Na Seção 7, é feita uma discussão sobre os resultados obtidos, bem como sobre as ameaças à validade do experimento. Por fim, na Seção 8, são apresentadas as conclusões.

2. Base Conceitual

Nesta Seção, são apresentados alguns conceitos necessários para o entendimento deste trabalho.

Comparamos o desempenho de quatro métodos utilizados para correspondência de textos, relacionados à tarefa de classificação de *fake news*, os quais são utilizados para mapear palavras e/ou textos em vetores de números reais, perfazendo um modelo de espaço vetorial, explicado na próxima seção. O primeiro conjunto inclui dois métodos amplamente utilizados, baseados em frequência de termos: TF-IDF e BM25. O segundo inclui dois métodos de incorporação semântica de palavras: *Word2Vec* e *Doc2Vec*.

2.1 Modelo de Espaço Vetorial

Proposto em (SALTON; WONG; CHUNG-SHU, 1975), o Modelo do Espaço Vetorial é uma abordagem que representa documentos de uma coleção como vetores em um espaço multidimensional. Os componentes do vetor que representam o documento são calculados com base na frequência dos termos e associados a pesos.

Em uma coleção com n documentos e m palavras, representam-se os documentos D_1, D_2, ..., D_n como vetores d_1, d_2, ..., d_n no espaço $\mathbb{R}_m$:

$$j = (w_{1j}, w_{2j}, \dots w_{mj}), para\ 1 \leq j \leq n$$

Onde $w_{1j} \dots w_{mj}$ são os pesos dos respectivos termos no documento D_j.

Nessa representação, o documento é considerado como um conjunto não-ordenado de palavras, denominado de *Bag of Words* (BOW). Assume-se, portanto, que a ordem relativa entre os termos no documento pode ser ignorada. Por exemplo, as sentenças *"eu dormi hoje"* e *"hoje eu dormi"* não apresentam diferenças. Por outro lado, as situações *"fui dar um passeio de patins"* e *"fui dar um patins de passeio"* possuem significados diferentes, embora sejam completamente idênticas no BOW.

2.1. Similaridade de Cossenos

Conforme vimos anteriormente, documentos em uma coleção de texto podem ser vistos como um conjunto de vetores de dimensão n. O grau de similaridade entre os documentos d_i e d_j pode ser dado pelo cosseno do ângulo formado pelos vetores correspondentes. Desta forma, a similaridade de cossenos é uma função baseada em palavras-chave (*tokens)* que mede a similaridade entre duas cadeias de caracteres, utilizando vetores no espaço dimensional reduzido (LI; HAN, 2013). É uma função útil para calcular a relevância de palavras em documentos, por meio do cálculo do cosseno entre dois vetores. O valor do cosseno permite descobrir a proximidade entre um termo e um documento (ou fração do documento), e entre documentos.

Dados dois vetores, um vetor *s* e outro vetor *k*, contendo *tokens (*palavras ou termos*)* de dois textos, associamos um peso *w* a cada token, de acordo com a frequência em que aparecem nos documentos. O cosseno do ângulo entre os dois vetores corresponde à similaridade entre os dois documentos: um tweet e uma notícia falsa, por exemplo.

Em outras palavras e do ponto de vista matemático, a similaridade do cosseno entre dois vetores é medida por meio do cálculo do cosseno do ângulo entre eles e é representada pela seguinte equação:

$$\cos \Theta \frac{\vec{a}\vec{b}}{||a||||b||}$$

Abstraindo como o cálculo é feito, até porque este é feito geralmente pelas máquinas, é importante entender que o valor do cosseno, que varia de 0 a 1, indica a similaridade entre os documentos. Quanto mais próximo a 1, mais similares são os documentos, uma vez que o ângulo Θ formado entre dois vetores iguais é igual a 0 e o $cos(0) = 1$. Consequentemente, quanto mais próximo de 0, menos similares eles são (AL-ANZI; ABUZEINA, 2017).

2.2. *Term Frequency–Inverse Document Frequency* (TF-IDF)

No Modelo de Espaço Vetorial, em se tratando dos pesos associados aos termos (*tokens* ou palavras), a importância de atribuir estes pesos é tão grande quanto a seleção de atributos ou, em outras palavras, quanto a seleção de palavras a serem consideradas. Uma forma simples de se atribuir pesos é usar a contagem *ti* de ocorrências dos termos no documento *d* (BUCKLEY, 1993).

$$d = (t1, t2, \ldots tn)$$

Entretanto, segundo (LUHN, 1958) as palavras não são iguais e algumas podem servir como discriminantes de documentos, enquanto outras, nem tanto. É possível que termos se sobressaiam uns sobre os outros e, se atribuirmos os pesos adequados, reforçamos esse comportamento. Existem várias formas para definir o peso de um termo. Uma das mais conhecidas e utilizadas é o TF-IDF (SALTON; BUCKLEY, 1988), onde:

- **TF** (*Term Frequency*): corresponde ao número de vezes que o termo aparece no documento. Os termos que são frequentemente mencionados em determinados documentos podem servir como discriminantes.
- **IDF** (*Inverse Document Frequency*): chamado de inverso da frequência do documento, pois desfavorece os termos presentes em muitos documentos. Quando os termos estão distribuídos em toda a coleção, mas não estão concentrados em poucos documentos, então, esses termos têm pouco ou nenhum poder de discriminação de relevância.

Desta forma, define-se o TF-IDF como o produto das partes TF × IDF. Apesar de possíveis variações, Salton e Buckley (1988) definem os componentes por:

$$TF = tf_{t,d}$$

$$IDF_t = log\left(\frac{N_D}{df_t}\right)$$

Onde:

$tf_{t,d}$ = número de ocorrências do termo *t* no documento *d;*

df_t = número de documentos que possuem o termo *t;*

N_D = total de documentos.

Considerando, por exemplo, o conjunto de documentos *(corpus)* composto pelas seguintes sentenças e cada sentença como um documento ou um *tweet*:

Documento 1: *"o gato viu um rato",*

Documento 2: *"o gato perseguiu o rato",*

Documento 3: *"o rato subiu o telhado".*

Este seria o histograma com a contagem de palavras do *corpus:*

['o': 5, 'rato': 3, 'gato': 2, 'viu': 1, 'um': 1, 'perseguiu': 1, 'subiu': 1, 'telhado': 1]

A Tabela 1 a seguir exemplifica o cálculo do TF-IDF para nosso exemplo. Cada linha representa um documento e cada célula o produto TF-IDF:

Tabela 1 - Exemplo de cálculo do TF-IDF

-	*o*	*rato*	*gato*	*viu*	*um*	*perseguiu*	*subiu*	*telhado*
Documento 1	0,00	0,00	0,18	0,48	0,48	0,00	0,00	0,00
Documento 2	0,00	0,00	0,18	0,00	0,00	0,00	0,00	0,00
Documento 3	0,00	0,00	0,00	0,00	0,00	0,00	0,48	0,48

Fonte: Elaborado pelo autor

Se a palavra aparece em todos os documentos, pela fórmula, a IDF será o Log de 1, ou seja, será zero, pois o Log de 1, em qualquer base, é zero, que multiplicado pela TF, também produzirá o valor zero. Em outras palavras, quanto mais a palavra aparece na coleção de documentos, menor será o seu peso, podendo ser zero, caso apareça em todos os documentos. Tomando como exemplo os casos das palavras *"gato"* e *"rato"*, no Documento 1, temos:

Para a palavra *"gato"*, o detalhamento do cálculo é o seguinte:

$TF = 1$, pois *"gato"* aparece 1 vez dentre as 5 palavras do Documento 1;

$IDF = \log\left(\frac{3}{2}\right) = \log(1{,}5) = 0{,}18$, pois *"gato"* aparece em 2 dos 3 documentos;

$TF\ x\ IDF = 1{,}0 * 0{,}18 = 0{,}18.$

Para a palavra *"rato"*, temos o seguinte cálculo:

$TF = 1$, pois *"rato"* aparece 1 vez dentre as 5 palavras do Documento 1;

$IDF = \log\left(\frac{3}{3}\right) = \log(1) = 0$, pois *"rato"* aparece em 3 dos 3 documentos;

$TF\ x\ IDF = 1 * 0 = 0.$

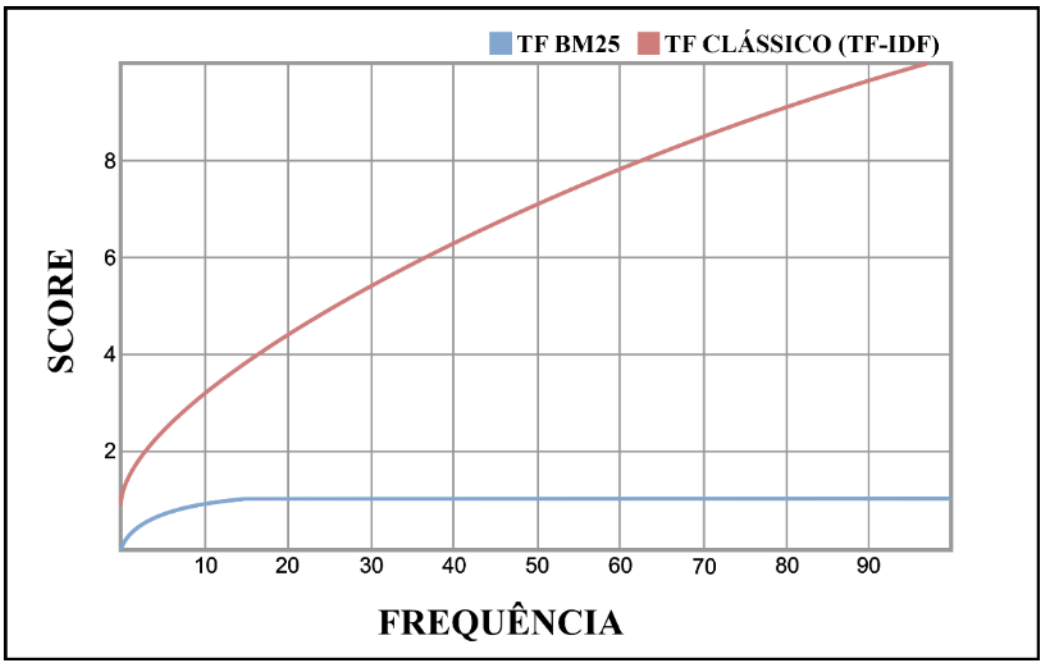
2.3. *Best Match 25* (BM25)

Baseado na teoria da probabilidade, o BM25 (ROBERTSON; ZARAGOZA, 2009) é uma função de classificação popular que também pode quantificar a importância da presença de cada palavra *(token)* para um determinado documento. Em tese, o BM25 representa uma

melhoria em relação ao TF-IDF, descrito na seção anterior. Por consequência, a formação do vetor é feita de forma semelhante.

No TF-IDF, o componente TF tende a ficar saturado muito rapidamente, especialmente para documentos curtos, como é possível observar na Figura 1. Um documento com 10 ocorrências de um termo é mais relevante do que 1 com somente uma ocorrência, mas não 10 vezes mais relevante. Idealmente, a relevância não deveria crescer proporcionalmente à frequência.

Figura 1 - Curva de saturação do componente TF



Fonte: (ELASTIC, 2020)

Para isto, o BM25 propõe um componente de TF mais penalizado:

$$tf(t,f) = \frac{((k+1) * f(t,d))}{(k * (1{,}0 - b + b * \frac{|d|}{|d|_{avg}} + f(t,d)}$$

Onde:

$|d|$ é o número de palavras no documento;

$|d|_{avg}$ é o número médio de palavras por documento;

$f(t,d))$ é o número de vezes que o termo *t* ocorre no documento *d*, comumente chamado de TF nos outros métodos;

k é um parâmetro ajustável que ajuda a determinar as características de saturação de frequência de termo. Por padrão, é definido como 1,2;

b é um parâmetro ajustável que, quanto maior, os efeitos do comprimento do documento em relação ao comprimento médio são mais amplificados. Por padrão, é definido como 0,75.

A pontuação final do BM25 pode ser calculada como:

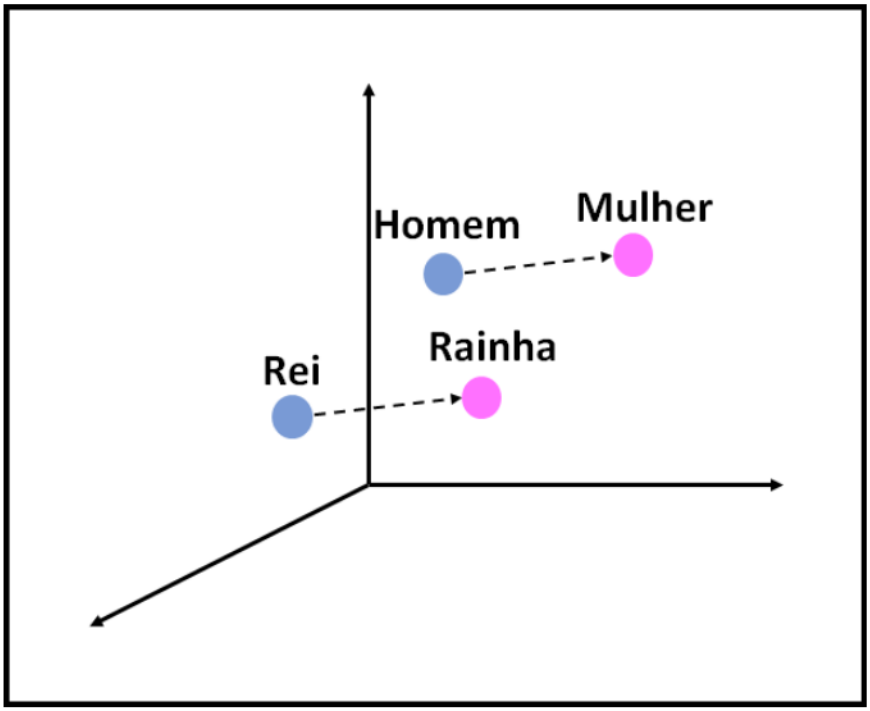
$$BM25(t,d) = \frac{((k+1) * f(t,d))}{(k * (1{,}0 - b + b * \frac{|d|}{|d|_{avg}} + f(t,d)} * IDF(t,D)$$

2.4. Word2Vec

O *Word2Vec* (MIKOLOV; SUTSKEVER *et al.*, 2013) é uma técnica de Processamento de Linguagem Natural (PLN) que utiliza um modelo de rede neural para aprender associações de palavras *(tokens)*, a partir do conjunto de documentos *(corpus)* usado para treinar o modelo sobre as associações. Com este modelo já treinado (existem modelos disponíveis já treinados no mercado), é possível detectar palavras sinônimas ou ainda sugerir palavras adicionais para uma frase. Para isto, cada palavra do *corpus* é transformada em um vetor numérico que a represente semanticamente.

Utilizando uma função matemática, como, por exemplo, a semelhança de cosseno entre os vetores, é possível indicar o nível de semelhança semântica entre as palavras e, de forma mais abrangente, entre documentos (RONG, 2014), como é possível observar na Figura 2, a seguir:

Figura 2 - Representação bidimensional simplificada para exemplificar a relação capturada entre palavras



Fonte: (CONTRATRES, 2020)

A ideia é que o espaço de incorporação consiga assimilar o contexto semântico. Assim, o modelo é capaz de interpretar que algo próximo de *rei*, por exemplo, poderá ser *rainha,* e que *homem* será mapeado para um vetor próximo de *mulher*. Também é possível encontrar similaridades entre pares de palavras, o exemplo mais encontrado na literatura é:

$$rei - homem + mulher = rainha$$

A similaridade a ser encontrada não significa que a expressão acima resulte no vetor *rainha*, mas sim que a palavra *rainha*, dentro do corpus usado para treinamento, é a mais próxima do vetor encontrado com a operação proposta. Desta forma, ao realizar esta operação em um ambiente de programação adequado, a palavra mais similar retornada será rainha. Lembrando sempre que tudo dependerá de como o algoritmo aprende, ou seja, um conjunto de textos sem lógica pode até colocar rainha longe de rei. Neste experimento, nosso *corpus*, ou seja, o conjunto de palavras, é formado a partir da base de notícias verificadas. Assim, cada palavra deste contexto possui um vetor que é usado no cálculo do nível de similaridade entre as notícias verificadas e *tweets*.

2.5. Doc2Vec

Proposto em (LE; MIKOLOV, 2014), o *Doc2Vec* é uma extensão do *Word2Vec* que aprende representações de frases de um documento, em um esquema de aprendizagem profunda supervisionada. Seu o objetivo é criar uma representação numérica de um documento, independentemente do seu comprimento, correlacionando rótulos e palavras, ao invés de palavras com outras palavras.

No método *Word2Vec,* não há necessidade de rotular as palavras, pois cada palavra tem seu próprio significado semântico no vocabulário. Porém, no caso do *Doc2Vec*, é necessário especificar quantas palavras ou frases transmitem um significado semântico, para que o algoritmo possa identificá-las como uma única entidade. Por esse motivo, especificam-se rótulos sentenciando documentos, dependendo do nível de significado semântico transmitido.

Se especificarmos um rótulo único para várias frases em um parágrafo, significa que todas as frases no parágrafo são necessárias para transmitir o significado. Por outro lado, se especificarmos rótulos variáveis para todas as sentenças de um parágrafo, significa que cada um transmite um significado semântico e eles podem ou não ter semelhança entre si. Em termos simples, rótulo significa significado semântico de alguma coisa. A ideia é chegar a um vetor que representa o significado de um documento, para associar documentos com rótulos.

2.6. Matriz de Confusão

Dentre as diversas formas de avaliar a capacidade de predição de um classificador para determinar a classe de vários registros, a matriz de confusão é considerada a mais simples (HAN; PEI; KAMBER, 2011) .

Para n classes, a matriz de confusão é uma tabela de dimensão $n \times n$. Para cada classificação possível, existe uma linha e coluna correspondente, ou seja, os valores das classificações serão distribuídos na matriz de acordo com os resultados, assim gerando a matriz de confusão para as classificações realizadas (PATRO; PATRA, 2014). As linhas correspondem às classificações corretas e as colunas representam as classificações realizadas pelo classificador (HAND; MANNILA; SMYTH, 2001) .

Quando existem apenas duas classes, uma é considerada como *positive* (no contexto desse trabalho, "Notícia Falsa") e a outra como *negative* ("Notícia Verdadeira") (HAND; MANNILA; SMYTH, 2001). Assim, podemos ter quatro resultados possíveis:

- *True Positive* (TP): uma instância de classe *positive* é classificada corretamente como *positive* (Notícia falsa, classificada corretamente como falsa);

- *True Negative* (TN): uma instância de classe *negative* é classificada corretamente como *negative* (Notícia verdadeira, classificada corretamente como verdadeira);

- *False Positive* (FP): uma instância de classe *negative* é classificada incorretamente como *positive* (Notícia verdadeira classificada como falsa);

- *False Negative* (FN): uma instância de classe *positive* é classificada incorretamente como *negative* (Notícia falsa classificada como verdadeira);

2.7. Métricas de Qualidade

Neste trabalho, foram utilizadas as métricas Acurácia, Precisão, Sensibilidade e Medida-F1 (CAELEN, 2017).

2.7.1 Acurácia

É o percentual de instâncias classificadas corretamente.

$$acuracia = \frac{TP + TN}{TP + TN + FP + FN}$$

2.7.2 Precisão

É a razão entre as instâncias classificadas como "verdadeiro positivo" e todas as instâncias classificadas como positivas.

$$precisao = \frac{TP}{TP + FP}$$

2.7.3 Sensibilidade

A sensibilidade, também conhecida como a taxa de verdadeiros positivos, recall ou cobertura real da amostragem positiva, é o percentual de instâncias que foram classificadas corretamente como positivas.

$$sensibilidade = \frac{TP}{TP + FN}$$

2.7.4 Medida-F1

Trata-se de uma média harmônica de duas medidas, pois combina a Precisão e a Sensibilidade, ponderando uniformemente.

$$medida - f1 = \frac{2 * precisao * sensibilidade}{precisao + sensibilidade}$$

3. Trabalhos Relacionados

As mídias sociais ganharam enorme popularidade em todo o mundo, tornando-se uma plataforma vital para a política. Muitos estudos, neste contexto, focam na análise de sentimentos (WANG; CAN *et al.*, 2012), ou na previsão de resultados de eventos políticos (TUMASJAN; SPRENGER *et al.*, 2010). Em paralelo, outros trabalhos focam na criação e circulação de *fake news* (FRIGGERI; ADAMIC *et al.*, 2014). Comparado a trabalhos existentes sobre detecção de *fake news* com foco em eventos sociais gerais ou eventos de emergência (JIN; CAO *et al.*, 2014), este artigo apresenta uma análise de *fake news*, no contexto de um evento político.

Uma boa parte dos trabalhos encontrados na literatura seguem o esquema tradicional de aprendizado de máquina supervisionado. Características do conteúdo do texto (CASTILLO; MENDOZA; POBLETE, 2011), clientes (MORRIS; COUNTS *et al.*, 2012), padrões de propagação (WU; YANG; ZHU, 2015), e conteúdo multimídia (JIN; CAO *et al.*, 2015) são extraídos para que um classificador de *fake news* aprenda com dados de treinamento rotulados (dados cuja classificação é previamente conhecida) como sendo ou não características de uma *fake*. Nesta mesma linha, alguns trabalhos recentes melhoraram os resultados deste tipo de classificação com métodos de otimização baseados em grafos (GUPTA; ZHAO; HAN, 2012). Contudo, embora as abordagens de aprendizado de máquina sejam muito eficazes, em algumas circunstâncias, existem algumas desvantagens. O processo de aprendizado supervisionado requer uma grande quantidade de dados para treinamento, os quais são difíceis de encontrar, além de serem, muitas vezes, computacionalmente caros.

Para superar os problemas de eficiência da aprendizagem supervisionada, em (ZHAO; RESNICK; MEI, 2015), foi proposto um método baseado em léxico para detectar *fake news* em *tweets*. Foram extraídas algumas palavras e frases para correspondência entre notícias verificadas e *tweets*, configurando um léxico relativamente pequeno, cujos resultados de detecção tendem a ter alta Precisão, mas baixa Sensibilidade.

Em outra dimensão e alternativa, em (JIN; CAO *et al.*, 2017), foi tratada a detecção de *fake news* como uma tarefa de correspondência de texto. Neste esquema, *tweets* e notícias previamente verificadas são comparadas por meio de métodos utilizados para verificar correspondência de textos, usando como medida a similaridade do cosseno, para calcular a distância entre um *tweet* e uma notícia. Além da classificação dos *tweets*, o resultado também mostra com qual notícia o mesmo está relacionado. Após avaliar quatro métodos, os autores obtiveram os seguintes resultados, em termos de Acurácia: TF-IDF - 79,50%, BM25 - 79,99%, *Doc2Vec* - 65,80% e *Word2Vec* - 55,70%. Neste experimento, surpreendentemente, ainda foi evidenciado que seguidores da candidata Hillary Clinton publicaram mais *fake news*, no entanto, os seguidores do candidato Donald Trump se mostraram mais ativos, no período mais próximo às eleições.

Analisando o contexto de polarização política, no qual as redes sociais digitais são um caminho pavimentado para o que passou a ser chamado de guerra híbrida (FERNANDES, 2016) , fomentada nas eleições para presidente dos EUA e do Brasil, é possível afirmar que o presente trabalho possui uma forte relação com o trabalho apresentado em (JIN; CAO *et al.*, 2017), em detrimento aos demais trabalhos relacionados. Desta forma, foi realizada uma replicação deste trabalho, dentro do contexto brasileiro, tendo como diferencial a utilização de uma abordagem experimental, com validação estatística da significância dos dados, o que permite uma replicação mais fiel dos procedimentos adotados, necessária para futuras metanálises dos resultados. Além disso, este artigo contribui com a consolidação da base de conhecimento já existente sobre os métodos de correspondência utilizados na detecção de *fake news* e terá sua metodologia resumida na próxima seção.

4. Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, um mapeamento sistemático da literatura, publicado em (AUTOR, 2020), tendo por finalidade encontrar o estado da arte das pesquisas sobre métodos de detecção de *fake news*. O mapeamento permitiu a identificação dos métodos mais utilizados para verificar correspondência de textos, no contexto das *fake news,* e a identificação de um trabalho que os avaliou, tendo como base a eleição presidencial americana de 2016 (JIN; CAO *et al.*, 2017). Pela similaridade com a nossa pesquisa, esse trabalho serviu de modelo e de controle para comparação dos resultados.

Do ponto de vista da classificação metodológica principal, este trabalho pode ser classificado como de laboratório e experimental, devido ao planejamento e a execução de um experimento controlado *"in vitro"*. Neste contexto, uma experimentação não é uma tarefa simples, pois envolve preparar, conduzir e analisar dados corretamente (WOHLIN; RUNESON *et al.*, 2012). Além disso, uma das principais vantagens da experimentação é o controle dos sujeitos, objetos e instrumentação, o que torna possível extrair conclusões mais gerais sobre o assunto investigado. Outras vantagens incluem a habilidade de realizar análises estatísticas, utilizando métodos de teste de hipóteses e oportunidades para replicação.

O experimento teve início com a coleta de dados de clientes do *Twitter* que publicaram informações no período eleitoral de 2018. Neste mesmo intervalo, foram obtidas de sites de *fact-checking*, notícias previamente checadas sobre as eleições e rotuladas como fatos (notícias verdadeiras) ou como *fake news*. Ambas as informações, *tweets* e notícias verificadas, as quais terão seus números detalhados na seção do experimento, passaram por um pré-processamento de texto. Nessa etapa, foram aplicadas técnicas de Processamento de Linguagem Natural, a fim de remover dos textos palavras desnecessárias (*stop-words*[1]). A irrelevância destas palavras depende do contexto a ser analisado. Como regra geral, pois, por exemplo, alguns contextos podem exigir a análise de numerais, são removidos preposições, artigos, conjunções, numerais e outros. Vejamos o exemplo utilizado para a explicação do TF-IDF:

Documento 1: *"o gato viu um rato"*,

Documento 2: *"o gato perseguiu o rato"*,

Documento 3: *"o rato subiu o telhado"*.

Ao realizar o pré-processamento, as palavras *"o"* e *"um"* seriam consideradas *stop-words*, ou seja, seus pesos seriam considerados irrelevantes em relação às demais palavras e estas seriam removidas imediatamente, antes dos cálculos dos *scores* de cada palavra.

Ainda no pré-processamento, *tweets* e notícias verificadas foram mapeados para vetores numéricos, utilizando cada um dos quatro métodos avaliados por este estudo, para que em seguida fosse aplicado o cálculo de similaridade entre eles.

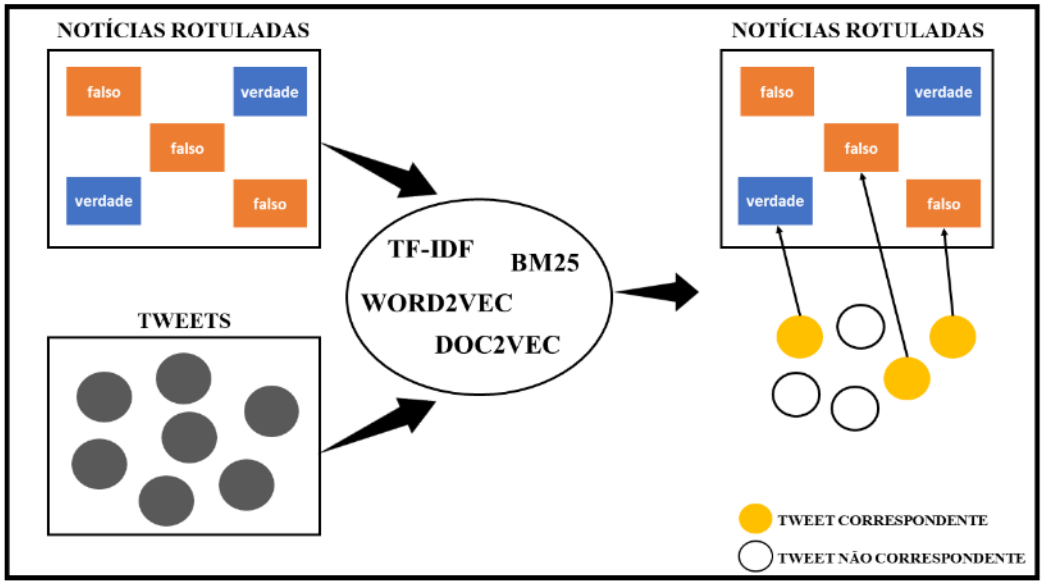
[1] Na computação, uma palavra vazia (ou stop-word, em inglês) é uma palavra que é removida antes ou após o processamento de um texto em linguagem natural.

Com os dados pré-processados, foi replicado o esquema de correspondência de texto usado em (JIN; CAO *et al.*, 2017), no qual é possível medir a similaridade entre dois documentos, ou entre um documento específico com todo o *corpus*, permitindo a identificação dos mais semelhantes, por meio da pontuação obtida no cálculo do cosseno do ângulo. Esta pontuação pode ser descrita como o nível de similaridade entre dois documentos. De posse dessa pontuação, foi definida uma linha de corte, ou limiar, para que fosse possível realizar a atribuição de um tweet a uma notícia falsa ou verdadeira.

Para determinação do limiar, foram feitas diversas e árduas avaliações com faixas de valores de limiares. Com limiares abaixo de 0,8, os resultados produzidos foram muito ruins, já para valores iguais ou superiores a 0,8, os resultados permaneceram quase em uma constante, configurando o valor adotado para este experimento.

Cada notícia existente na base coletada possui um rótulo, verdadeira ou falsa. Utilizando a similaridade do cosseno, cada *tweet* é comparado com cada notícia existente na base. Ao final, a maior pontuação de cada comparação é atribuída como o nível de similaridade. Caso este valor seja maior ou igual ao limiar, é possível dizer sobre qual notícia o *tweet* se refere. Além disso, uma vez que as notícias são previamente rotuladas, é possível classificar o *tweet* como verdadeiro ou falso, de acordo com a notícia a qual ele corresponde. O esquema de correspondência de texto descrito pode ser observado na Figura 3.

Figura 3 - Modelo de correspondência entre documentos utilizados



Fonte: Elaboração do autor

Para a obtenção das métricas a serem avaliadas, utilizou-se uma adaptação da abordagem *10-Fold Cross-validation* (HASTIE; TIBSHIRANI; FRIEDMAN, 2011). Em nosso contexto, com a base de notícias já rotulada e levando em consideração o caráter único de cada notícia e *tweets* que se aproximam destas, ou seja, sem considerar características gerais de uma notícia *fake* e sem separar partes da base de treinamento (notícias rotuladas) para testes, a base de *tweets* foi dividida em 10 partes e cada método foi avaliado em todas as partes, perfazendo sempre, para cada métrica de qualidade, 10 medidas calculadas para cada método.

Em se tratando do cálculo da estimativa geral do número de *fake news* por seguidores de cada candidato, foi calculada e utilizada como base uma amostra para população infinita, perfazendo 801 *tweets* checados manualmente, a fim de validar a similaridade entre notícias e *tweets,* no esquema de correspondência utilizado por esta pesquisa. O detalhamento desta amostragem será descrito na próxima seção.

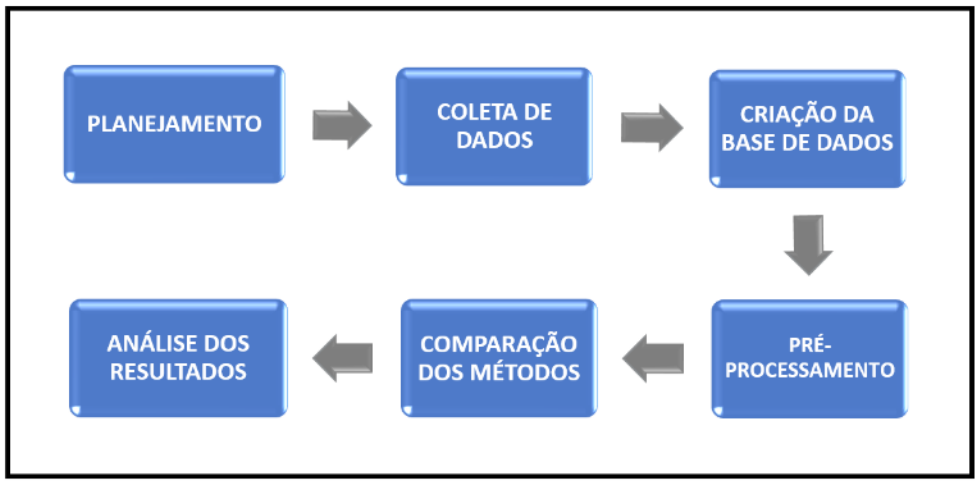
Finalmente, para auxiliar nos cálculos e verificar possíveis diferenças significativas na eficácia dos algoritmos, foi utilizado a ferramenta de análise de dados SPSS (*Statistical Package for Social Science*) (SPSS, 2020), com a qual foram aplicadas técnicas estatísticas básicas e avançadas. O SPSS é um software estatístico internacionalmente utilizado há muitas décadas, desde suas versões para computadores de grande porte (MUNDSTOCK; GUIMARÃES *et al.*, 2006).

Em resumo, o experimento pode ser dividido em quatro etapas principais: planejamento; operação de limpeza dos dados, coleta e geração do conjunto de dados; comparação de métodos; e, finalmente, a análise dos resultados. O experimento em questão é detalhado nas próximas seções.

5. Definição e Planejamento do Experimento

Nesta e nas duas próximas seções, este trabalho é apresentado como um processo experimental. O mesmo segue as diretrizes apresentadas em (OLIVEIRA; COLAÇO JÚNIOR, 2018). A Figura 4 ilustra as etapas do trabalho, esta Seção irá focar na etapa 5.2, o planejamento do experimento.

Figura 4 - Etapas do trabalho



Fonte: Elaboração do autor

5.1. Definição do Objetivo

O objetivo deste trabalho é fazer uma análise experimental dos principais métodos utilizados para verificar correspondência de textos encontrados na literatura, para detecção de *fake news*, avaliando e validando o que melhor se adéqua ao contexto de *fake news* no processo eleitoral brasileiro.

Utilizando o modelo GQM (*Goal Question Metric*) (BASILI; WEISS, 1983), foi possível formalizar o objetivo deste estudo da seguinte maneira: **Analisar**, por meio de experimento controlado, os principais métodos utilizados para verificar correspondência de textos aplicados ao contexto de *fake news*, **com a finalidade de** avaliá-los (contra resultados de trabalhos anteriores realizados para eleição norte-americana), **com respeito à** Acurácia, Precisão, Sensibilidade e Medida-F1, **do ponto de vista de** cidadãos, pesquisadores e profissionais de Ciência de Dados, **no contexto** das *fake news* sobre as eleições presidenciais brasileiras de 2018.

5.2. Planejamento

5.2.1 Seleção de Contexto

O experimento foi realizado *"in vitro"*, considerando dados de clientes do *Twitter* publicados no período eleitoral de 2018. As ações por parte de partidos e candidatos, tais como registro de candidatura, convenções ou filiação, que fazem parte do período eleitoral, iniciam-se bem antes do pleito. Desta forma, visando obter a maior quantidade de *tweets* relacionados

a notícias nesse contexto, foram coletadas publicações do período entre junho, quando já havia movimentações políticas, e dezembro de 2018, uma vez que, mesmo após o fim do pleito, a movimentação nas redes sociais ainda era alta.

5.2.2 Formulação de Hipóteses

Para guiar o estudo, foi elaborada a seguinte questão principal de pesquisa, cuja resposta visa cumprir o objetivo do trabalho. No contexto da detecção de *fake news* no *Twitter*, entre os métodos selecionados, qual o melhor em termos das métricas avaliadas?

Para avaliar esta questão, foram utilizadas quatro métricas: Acurácia, Precisão, Sensibilidade e Medida-F1.

Sendo assim, com os objetivos e métricas definidas, serão consideradas as hipóteses a seguir (**para cada métrica**). A avaliação a ser feita pretenderá rejeitar ou não rejeitar a hipótese nula (H_0):

- H_0: Os métodos$_{(1,\ 2\ldots n)}$ possuem médias iguais para a métrica.
$\mu 1(métrica) = \mu 2(métrica) \ldots = \mu n(métrica)$;

- H_1: Os métodos$_{(1,\ 2\ldots n)}$ possuem médias diferentes para a métrica.
$\mu 1(métrica) \neq \mu 2(métrica) \ldots \neq \mu n(métrica)$;

5.2.3 Seleção de Participantes e Objetos

Na coleta dos dados, os seguidores de cada candidato precisaram cumprir algumas premissas. Primeiramente, não seguir ambos os candidatos, dessa forma, tentou-se aproximar-se de perfis de eleitores reais. Além disso, sua configuração de privacidade deveria estar como pública, permitindo assim o acesso e visualização dos *tweets* e informações do perfil.

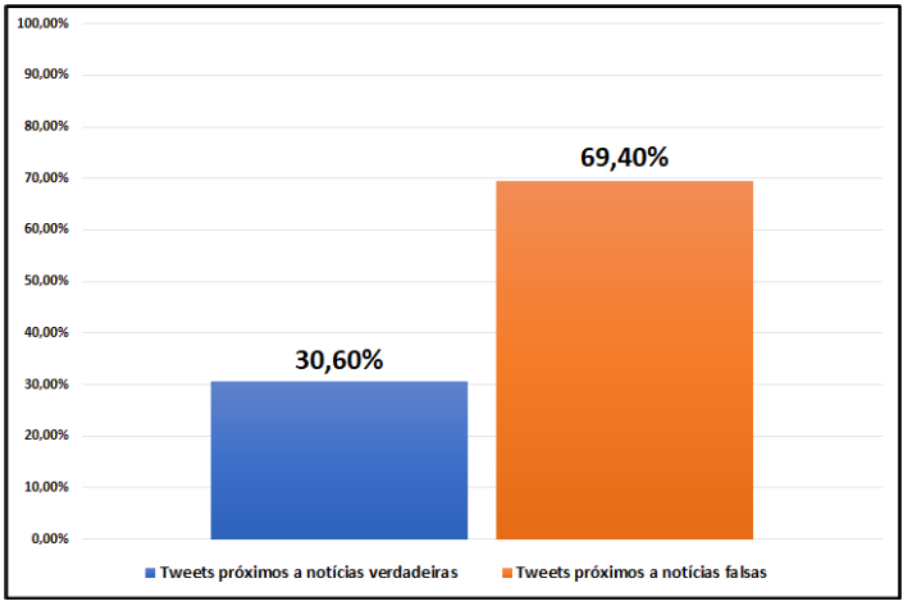
Definidos o período de coleta e as regras para os perfis dos clientes, coletou-se *tweets* de seguidores de ambos os candidatos. A fim de agilizar o processo de obtenção dos dados, a coleta foi realizada em paralelo, ou seja, foi possível coletar os dados dos seguidores de cada um dos candidatos, simultaneamente, e, em seguida, foram consolidadas as duas bases de dados distintas.

Com um caráter surpreendentemente balanceado, foram coletados 1.155.078 (49,99%) perfis de clientes que seguiam o candidato Jair Bolsonaro, até então membro do Partido Social

Liberal (PSL), e 1.155.140 (50,01%) perfis de clientes seguidores do candidato do PT, Fernando Haddad, somando, no total, 2.310.218 perfis.

Considerando o limiar de *0,8* para a medida de similaridade do cosseno, ou seja, selecionando apenas os *tweets* que se aproximaram de uma notícia checada, verdadeira ou falsa, com um nível de similaridade maior ou igual a 0,8 (cosseno do ângulo entre o *tweet* e a notícia checada maior ou igual a 0,8), obteve-se um conjunto de dados com 2847 *tweets*. Desse total, 1.964 *tweets* se aproximavam de *fake news* e 883 eram sobre fatos verídicos, como mostra a Figura 5.

Figura 5 - Distribuição de tweets por rótulo



Fonte: Elaboração do autor

Estes resultados colaboram com os resultados apresentados em (VOSOUGHI; ROY; ARAL, 2018). Segundo esta pesquisa, uma *fake news* tem 70% mais chances de ser compartilhada do que uma notícia verdadeira. Além disso, informações falsas ganham espaço na internet de forma mais rápida, mais profunda e com mais abrangência que as verdadeiras, pois tendem a ser mais impactantes ou inéditas, o que acaba atraindo as pessoas, que movidas pela sensação de privilégio ou ineditismo, divulgam rapidamente a informação.

Sobre quem publicou, 1.679 *tweets* vieram de seguidores de Jair Bolsonaro, contra 1.168 oriundos de seguidores de Fernando Haddad. No grupo de seguidores de Jair Bolsonaro, dos 1.679 *tweets*, 1326 se aproximaram de *fake news* e 353 eram mais próximos de notícias

verdadeiras. Entre os seguidores de Fernando Haddad, dos 1.168 *tweets*, 794 se aproximavam de *fake news* e 374 se aproximam de notícias verdadeiras.

Como não era possível afirmar, imediatamente, que todos o tweets próximos a *fake news* ou verdadeiras eram, respectivamente, verdadeiros positivos ou verdadeiros negativos, para que fosse gerada uma visão geral de *fake news* por seguidores, bem como para a avaliação das métricas e validação da classificação feita com base no nível de similaridade, realizou-se uma árdua checagem manual de uma amostra dessas informações. Deste modo, foi calculada e selecionada uma amostra de 801 *tweets*.

Para o cálculo da amostra (SEWARD; DOANE, 2014), foi considerada toda a população de seguidores e *tweets*. Consideramos uma amostra de *tweets*, com margem de erro de 3,46% e confiabilidade de 95%, para a população de 2.310.218 seguidores e 1.845.603 de *tweets*, perfazendo 801 *tweets*, os quais foram divididos dentro da proporção geral de seguidores, surpreendentemente balanceada, aproximadamente 49,99% (1.155.078) de seguidores de Bolsonaro e 50,01% (1.155.140) de seguidores de Haddad. A reflexão aproximada desta proporção na amostra foi consistida por 401 *tweets* de seguidores de Haddad e 400 *tweets* de seguidores de Bolsonaro. Vale ressaltar que a amostra, em verdade, foi arredondada e um pouco maior do que o cálculo de uma amostra para uma população infinita, a qual, considerando a mesma confiabilidade e uma margem de erro de 3,5%, seria de 784 tweets.

Calculada a amostra, do ponto de vista da seleção, esta foi feita de forma aleatória, por meio de uma função de randomização programada no banco de dados, com a qual foram sendo sorteados números de linhas da base de dados de *tweets* e estes foram sendo recuperados e consultados. Em seguida, verificava-se manualmente se a notícia era falsa ou verdadeira, por meio da comparação a olho nu com as notícias já verificadas pelas agências de checagem de fatos.

5.2.4 Variáveis independentes

As variáveis independentes referem-se à entrada do processo de experimentação, ou seja, representa a causa que afeta o resultado do experimento (TRAVASSOS; GUROV; AMARAL, 2002). Para este trabalho, foram consideradas como variáveis independentes os conjuntos de notícias checadas, os *tweets* coletados, o limiar usado para definir a classe de um *tweet* e os métodos para mapeamento de textos e palavras.

5.2.5 Variáveis dependentes

As variáveis dependentes abordadas no experimento foram as classificações, tendo como derivação as medidas de interesse objetivas para auxiliar na identificação da qualidade destas: Acurácia, Precisão, Sensibilidade e Medida-F1.

5.2.6 Projeto do Experimento

Uma das métricas utilizadas neste trabalho foi a Acurácia, a qual exige o balanceamento dos dados das classes. Uma vez que os dados coletados já estão balanceados (MACHADO, 2007), não foi necessário planejar a adoção de um método de balanceamento.

Além disso, conforme descrito na metodologia, neste experimento, foi utilizada uma adaptação da abordagem *10-Fold Cross-Validation* (HASTIE; TIBSHIRANI; FRIEDMAN, 2011). O conjunto de *tweets* foi dividido em 10 partes, sendo obtidos 10 valores de cada métrica, para cada método avaliado. Posteriormente, foram calculadas as médias das métricas, para validação estatística.

5.2.7 Instrumentação

O processo de instrumentação consistiu na configuração do ambiente para a realização do experimento controlado. Os materiais/recursos utilizados foram: biblioteca *Scikit-learn* de aprendizado de máquina de código aberto, para a linguagem de programação Python (PEDREGOSA; VAROQUAUX *et al.*, 2011), *Twitter API (Application Programming Interface,)* (MAKICE, 2009), usada para extrair dados fornecidos pela rede social, SPSS (SPSS, 2020), *Amazon Web Services* (AWS) (AMAZON, 2019) e um computador com Intel(R) Core(TM) i5-5200 CPU a 2,20GHz, 12GB de RAM - 64 bits. A preparação do ambiente de testes foi feita baixando e instalando todas as bibliotecas mencionadas.

6. Operação do Experimento

6.1 Preparação

Os dados utilizados neste experimento foram obtidos a partir de diferentes fontes de dados, consequentemente, deram origem a duas bases, uma com notícias checadas e outra com *tweets* dos seguidores de ambos os candidatos.

A base com notícias possui dados coletados de três diferentes sites de agência de checagem de fatos:

- Aos Fatos (FATOS, 2018);
- Agência Pública (PÚBLICA, 2018);
- Lupa (LUPA, 2019).

Para a obtenção dessas informações, foi utilizada uma técnica chamada *Web Scraping,* com a qual é possível extrair informações relevantes de um determinado site. Desta forma, foi desenvolvido um programa na linguagem *Python*, o qual coletou 460 notícias sobre as eleições brasileiras de 2018. Deste conjunto de notícias, 238 (51,73%) estavam rotuladas como falsas e 222 (48,27%) como verdadeiras, perfazendo nossa base de notícias rotuladas.

Na coleta dos *tweets*, utilizou-se a API (*Application Programming Interface*) disponibilizada pela rede social. Todos os dias, milhares de requisições são feitas à plataforma de desenvolvedores do *Twitter*. Para ajudar a gerenciar este volume, são impostos limites às solicitações que podem ser feitas. Desta forma, visando agilizar o processo de coleta, foram configuradas duas máquinas virtuais do tipo EC2 (*Elastic Compute Cloud*), na *Amazon,* para realizar coleta de forma simultânea. Por meio de programas escritos na linguagem *Python*, as máquinas virtuais coletaram dados dos clientes seguidores de cada um dos candidatos, 24h por dia, 7 dias por semana. A lógica aplicada a este algoritmo já contemplava pausas necessárias, após uma certa quantidade de requisições à API do *Twitter*. Em seguida, informações como *hashtags*, menções, mídias e *tweets* foram coletadas e armazenadas em instâncias de um banco de dados orientado a documentos (MONGODB, 2020).

A coleta e armazenamento dos *tweets* e das notícias contemplam as etapas 2 e 3 da Figura 4. Antes de executar o mapeamento realizado pelos métodos aqui avaliados, foi realizada uma etapa de pré-processamento de texto. Nas notícias, foram utilizadas as bibliotecas *NLTK* (BIRD, 2020), para a remoção de *stop-words,* e *Num2words* (OGAWA, 2020), para a conversão de números em números por extenso. No pré-processamento dos *tweets,* além das bibliotecas citadas anteriormente, também foi utilizada a biblioteca *Tweet Pre-Processor* (ÖZCAN, 2020), a qual já possui funções nativas para a remoção de atributos como *hashtags*, *emojis* e *retweets*. Desta forma, informações desnecessárias ao cálculo de similaridade foram removidas. A seguir, é possível observar dois exemplos de *fake news* checadas e de dois *tweets* antes e após essa etapa:

***Fake news* Checadas:**

- *Antes:* Empresa contratada pelo TSE para apuração dos votos tem ligação com o PT

- *Depois:* empresa contratada tse apuracao votos ligacao pt
- *Antes:* Terrorista é Bolsonaro, que foi processado e expulso do exército
- *Depois:* terrorista bolsonaro foi processado expulso exercito

***Tweets*:**

- *Antes:* Vamos abrir os olhos, fraude nas apurações...Empresa contratada pelo TSE tem ligação direta com PT...#BrasilComBolsonaro
- *Depois:* vamos abrir olhos fraude apuraçoes empresa contratada tse ligacao direta
- *Antes:* Imagina ter que lembrar pra alguém que Bolsonaro era terrorista e foi expulso do exército?
- *Depois:* imagina lembrar alguem bolsonaro terrorista foi expulso exercito

6.2 Execução

Consistiu na realização da classificação dos *tweets,* de acordo com as notícias rotuladas, conforme planejado na Subseção 5.2.6, para cada método selecionado, utilizando o dicionário discutido na Subseção 5.23. Etapa 4 da Figura 4.

6.3 Validação dos Dados

Para análise, interpretação e validação - etapa 5 da Figura 4, foram utilizados seis tipos de testes estatísticos: *Anova, Friedman, Levene, Shapiro-Wilk*, *Tukey e Wilcoxon*.

O Teste *Anova* foi utilizado por ser necessário comparar mais de dois grupos de valores. Como este teste possui os pressupostos de que a distribuição deve ser Normal e de que haja homocedasticidade entre os tratamentos (variâncias homogêneas) (FIELD, 2009), foi utilizado o teste *Shapiro-Wilk* (SHAPIRO; WILK, 1965), para o teste de Normalidade, e o teste de *Levene* (LEVENE, 1960), para o teste de homocedasticidade.

O Teste *Anova* evidencia que ao menos um método se diferencia dos demais, mas não é possível afirmar qual é o mais discrepante. Para isso, foi utilizado o teste *Tukey*, que, segundo Anjos (2009), permite testar qualquer contraste, sempre, entre duas médias de tratamentos, sendo possível verificar quais são estatisticamente iguais ou diferentes.

Os testes de *Friedman* (FRIEDMAN, 1937) *e Wilcoxon* (WILCOXON, 1945) foram utilizados para a comparação das medianas da Medida-F1, uma vez que, para esta métrica, o resultado do teste de normalidade indicou uma distribuição de dados não-normal.

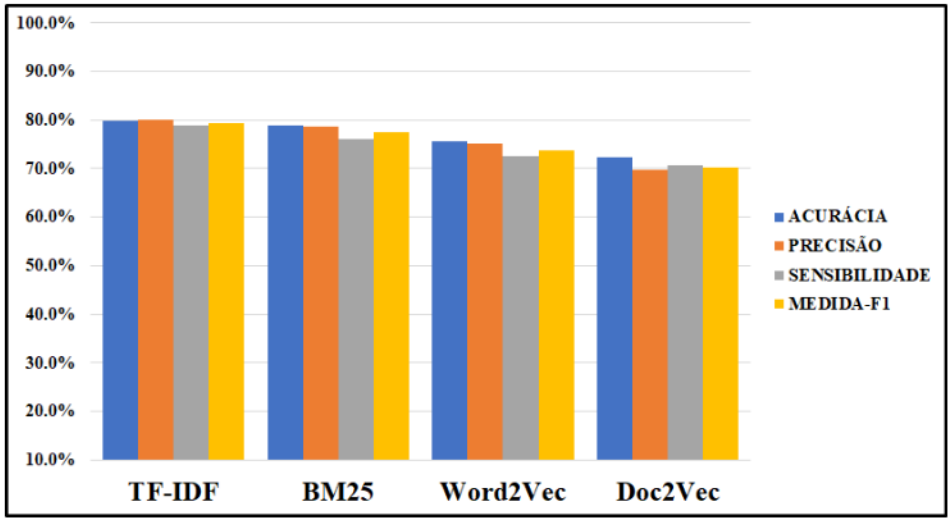
Todos os testes estatísticos foram feitos utilizando a ferramenta SPSS – IBM (SPSS, 2020).

7. Resultados

7.1 Análise e Interpretação de Dados

Após a execução do experimento, foram obtidos os resultados das classificações alcançadas por cada um dos quatro métodos utilizados para verificar correspondência de textos avaliados. Na Tabela 2 e na Figura 6, são apresentadas as médias das métricas.

Figura 6 - Comparativo das métricas entre os métodos



Fonte: Elaboração do autor

Tabela 2 - Comparativo das métricas para os métodos

Método	*Acurácia*	*Precisão*	*Sensibilidade*	*Medida-F1*
TF-IDF	**79,86%**	**79,97%**	**78,97%**	**79,47%**
BM25	79,00%	78,76%	76,05%	77,38%
Word2Vec	75,69%	75,04%	72,65%	73,83%
Doc2Vec	72,39%	69,85%	70,77%	70,31%

Fonte: Elaboração do autor

Estes resultados foram utilizados para responder à questão de pesquisa definida na Seção 5.2. Como é perceptível, os métodos obtiveram médias de Acurácias distintas e o método TF-IDF obteve a maior média, seguido pelo BM25. Assim como no trabalho de (JIN; CAO *et al.*, 2017), *Doc2Vec* e *Word2Vec* obtiveram resultados um pouco abaixo dos demais. Todavia, não é possível fazer essas afirmações sem evidências estatísticas suficientemente conclusivas.

Como já mencionado anteriormente, o teste *Anova* foi aplicado para validar a hipótese, e, por possuir os pressupostos da normalidade e da homocedasticidade, primeiramente, foi feito o teste *Shapiro-Wilk e*, em seguida, o teste de *Levene*. Para o caso no qual o teste de normalidade não foi satisfatório, foi aplicado o teste de *Friedman*, descrito posteriormente, como uma alternativa não paramétrica ao teste *Anova.*

Definiu-se um nível de significância (α) de 0,05 em todo o experimento. Ao aplicar o teste de *Shapiro-Wilk*, para análise da normalidade da distribuição dos dados, foram obtidos os *p-values* apresentados na Tabela 3, na qual, observa-se 3 valores acima do nível de significância adotado, concluindo-se que estas distribuições são normais, com exceção da distribuição da Medida-F1, para o método TF-IDF.

Tabela 3 - Resultado do Teste de Shapiro-Wilk, para análise da normalidade dos dados

Método	*Acurácia*	*Medida-F1*
TF-IDF	0,172	0,031
BM25	**0,751**	0,297
Word2Vec	0,256	**0,954**
Doc2Vec	0,241	0,371

Fonte: Elaboração do autor

Em seguida, foi realizado o teste de *Levene,* para a acurácia, pois, neste caso, não houve rejeição da normalidade para nenhum método. O resultado obtido é apresentado na Tabela 4. Como pode ser observado, o *p-value* obtido é maior que o nível de significância adotado, validando o pressuposto da homogeneidade de variâncias entre os métodos.

Tabela 4 - p-values dos Testes de Levene, Anova e Friedman

Métricas	*Levene*	*Anova*	*Friedman*
Acurácia	0,176	< 0,001	-
Medida-F1	-	-	< 0,001

Fonte: Elaboração do autor

Uma vez que os pressupostos foram atendidos, foi possível aplicar o teste Anova para a Acurácia, com o qual se verificou um *p-value* fortemente menor que o nível de significância adotado, como pôde ser observado na Tabela 4. Desta forma, foi possível confirmar a evidência da diferença entre as médias, ou seja, a hipótese $H_{(0)}$, de que os métodos possuem a mesma Acurácia, foi rejeitada, dentro do contexto do experimento realizado.

Sendo assim, com o teste *Anova*, foi evidenciado que ao menos um método se diferencia dos demais, porém, não é possível afirmar qual é o mais discrepante. Para isso, foi utilizado o teste *Tukey* para uma análise posterior (*post-hoc)*. A Tabela 5, a seguir, apresenta as médias das Acurácias dos métodos agrupados, formando três grupos homogêneos. É possível observar que a maior média foi a do TF-IDF, 79,86%, contudo, do ponto de vista estatístico, conforme grupo apresentado na tabela, similar à média do BM25. O *Word2Vec* e *Doc2Vec* apresentaram as médias mais baixas, com 75,69% e 72,39%. Estes resultados confirmam evidências anteriores encontradas nos trabalhos descritos em (MÁRQUEZ-VERA; MORALES; SOTO, 2013) e (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009).

Tabela 5 - Valores obtidos pelo Teste de Tukey para Acurácia

Método	*Subconjunto para alfa = 0.5*		
	1	**2**	**3**
Doc2Vec	72,39%		
Word2Vec		75,69%	
BM25			79,00%
TF-IDF			79,86%
SIG.	1	1	0,887

Fonte: Elaboração do autor

Com relação às métricas de precisão e sensibilidade, os dados não serão apresentados, uma vez que a Medida-F1 harmoniza estas métricas. Para esta medida, foi utilizado o teste *Friedman*, uma vez que o resultado do teste de normalidade indicou uma distribuição não-normal e este teste é uma alternativa ao *Anova*.

Com a aplicação do teste de *Friedman* para a medida-F1, verificou-se um *p-value* fortemente menor que o nível de significância adotado, como pôde ser observado na Tabela 4. Desta forma, foi possível confirmar a evidência da diferença entre as medianas, ou seja, a hipótese $H_{(0)}$, de que os métodos possuem a mesma medida-F1, foi rejeitada, dentro do contexto do experimento realizado.

Desta forma, similar ao caso da Acurácia, foi evidenciado que ao menos um método se diferencia dos demais, porém, não é possível afirmar qual é o mais discrepante. Para isso, foi utilizado o teste de *Wilcoxon,* uma alternativa não paramétrica ao teste de *Tukey*. A Tabela 6 apresenta o resultado deste teste para a Medida-F1, evidenciando que, após uma análise *"post-hoc"*, ou posterior, aplicando a correção de Bonferroni ($\alpha = \alpha / 6$), encontramos a seguinte ordem relacionada aos métodos: TF-IDF > *Word2Vec*, *Doc2Vec*, no entanto, também não houve significância estatística para não rejeitar que o TF-IDF foi superior ao BM25. Nesta análise posterior, cada método foi comparado aos outros, em avaliações dois a dois. A significância estatística é verificada nas linhas em que o *p-value* é menor que 0,05. Em outras palavras, cada linha da Tabela 6 testa a hipótese nula de que as distribuições da Medida-F1 do Método 1 e da Medida-F1 do Método 2 são iguais. O nível de significância continua sendo 0,05.

Tabela 6 - Valores obtidos pelo Teste de Wilcoxon, dois a dois

Comparações Dois a Dois		
Método 1-Método 2	*p-value*	*p-value* ajustado[2]
Doc2Vec-*Word2Vec*	0,083	0,500
Doc2Vec-BM25	< 0,001	0,003
Doc2Vec-TF-IDF	< 0,001	0,000
Word2Vec-BM25	0,083	0,500
Word2Vec-TF-IDF	< 0,001	0,003

[2] Os valores de significância foram ajustados pela correção de Bonferroni para vários testes.

BM25-TF-IDF	0,083	0,500

Fonte: Elaboração do autor

Na Tabela 7, é possível observar os resultados, em termos de Acurácia e Medida-F1, obtidos por este trabalho, em comparação com o trabalho apresentado em (JIN; CAO *et al.*, 2017). Neste experimento, TF-IDF e BM25 obtiveram resultados similares. Considerando a aplicação dos testes estatísticos feitos por este estudo, não foi possível evidenciar significância estatística para diferença entre seus resultados. No trabalho apresentado em (JIN; CAO *et al.*, 2017), o BM25 superou o TF-IDF em 0,49%, em termos de Acurácia, e 6,2%, em termos de Medida-F1. Mesmo assim, considerando que não há evidências de análise de significância estatística por parte do trabalho replicado, o empate técnico entre os dois métodos também pode ter ocorrido.

Tabela 7 - Comparativo de resultados entre este trabalho e o estudo replicado

-	**Trabalho Atual**		**Trabalho Replicado**	
Método	*Acurácia*	*Medida-F1*	*Acurácia*	*Medida-F1*
TF-IDF	**79,86%**	**79,47%**	79,50%	75,80%
BM25	79,00%	77,38%	**79,99%**	**82,00%**
Word2Vec	75,69%	73,83%	55,70%	76,40%
Doc2Vec	72,39%	70,31%	65,80%	74,50%

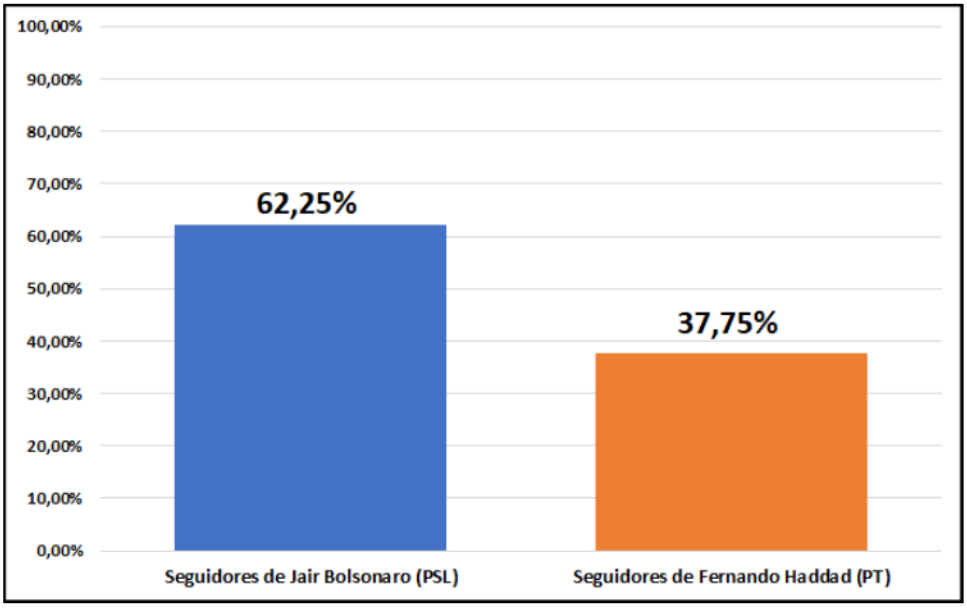
Fonte: Elaboração do autor

Ao analisar o desempenho do *Word2Vec* e do *Doc2Vec*, nota-se uma melhora em seus resultados para este experimento. No entanto, em ambos os trabalhos, seus resultados foram piores do que os do TF-IDF e do BM25, considerados métodos tradicionais. Desconsiderando questões de desempenho, novas pesquisas precisam analisar se os resultados destes métodos melhoram, combinando-os com um modelo de aprendizagem profunda (*Deep Learning*) (SANTOS; COLAÇO JÚNIOR *et al.*, 2020). Além disso, considerando que a área de *fake news* possui suas particularidades e idiossincrasias, o uso da incorporação de palavras (*Word Embeddings*[3]) evidencia inicialmente a necessidade de adaptações para este novo contexto.

[3] Dado um texto, são as representações das palavras em vetores de números reais, os quais contêm algum conhecimento das informações de posicionamento entre as palavras (vide seção de base conceitual).

Partindo para uma análise mais próxima ao "negócio" eleições, os resultados demonstrados na Figura 7 fazem uma comparação da quantidade de *tweets* relacionados a *fake news* entre seguidores de cada um dos candidatos. Conforme descrito anteriormente, foi calculada e utilizada uma amostra, com margem de erro de aproximadamente 3,5%, e um nível de confiabilidade de 95%. Os seguidores do candidato Jair Bolsonaro foram responsáveis por 62,25% do total de *tweets* sobre *fake news*, contra 37,75% publicados por seguidores de Fernando Haddad. Neste sentido, mesmo com uma maior quantidade de *fake news* publicadas por parte dos seguidores do candidato do PSL, existiu disseminação de *fake news* por seguidores de ambos os candidatos.

Figura 7 - Porcentagem de tweets sobre fake news por seguidores de cada candidato

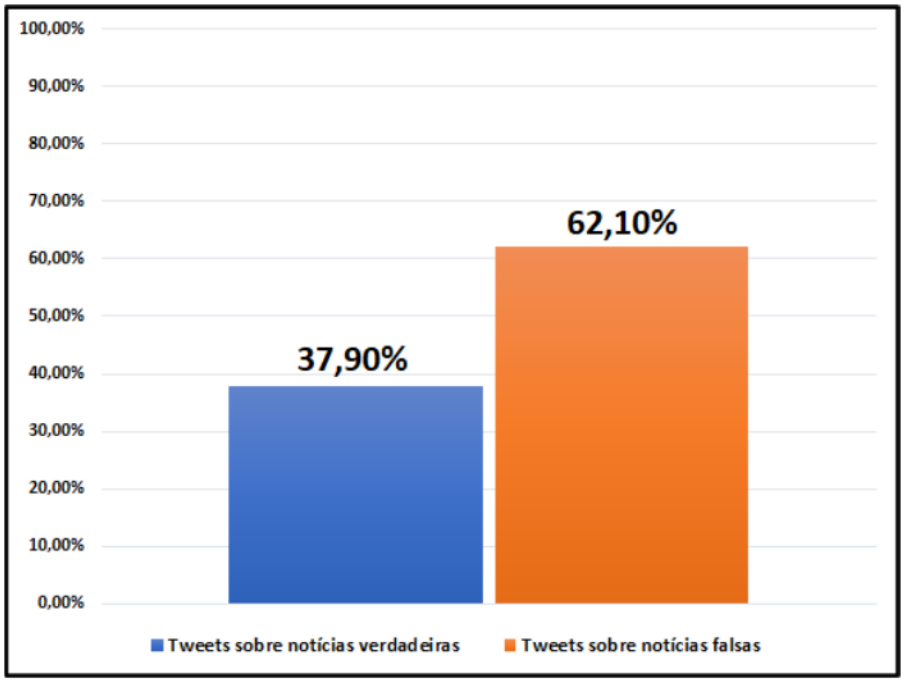


Fonte: Elaboração do autor

No pleito norte-americano, também houve disseminação de *fake news* por ambos os lados, com seguidores do candidato eleito, Donald Trump, surpreendentemente, tendo publicado 18% a menos de *tweets* sobre *fake news* do que os seguidores da candidata Hillary Clinton. No entanto, curiosamente, uma análise temporal mostrou que, no período mais próximo às eleições, os seguidores de Trump foram mais ativos no *Twitter* (JIN; CAO *et al.*, 2017).

Analisando os resultados de cada candidato separadamente, nota-se que, entre os seguidores do candidato Fernando Haddad (PT), seguindo a tendência discutida anteriormente, houve mais *tweets* sobre *fake news*, como é possível observar na Figura 8.

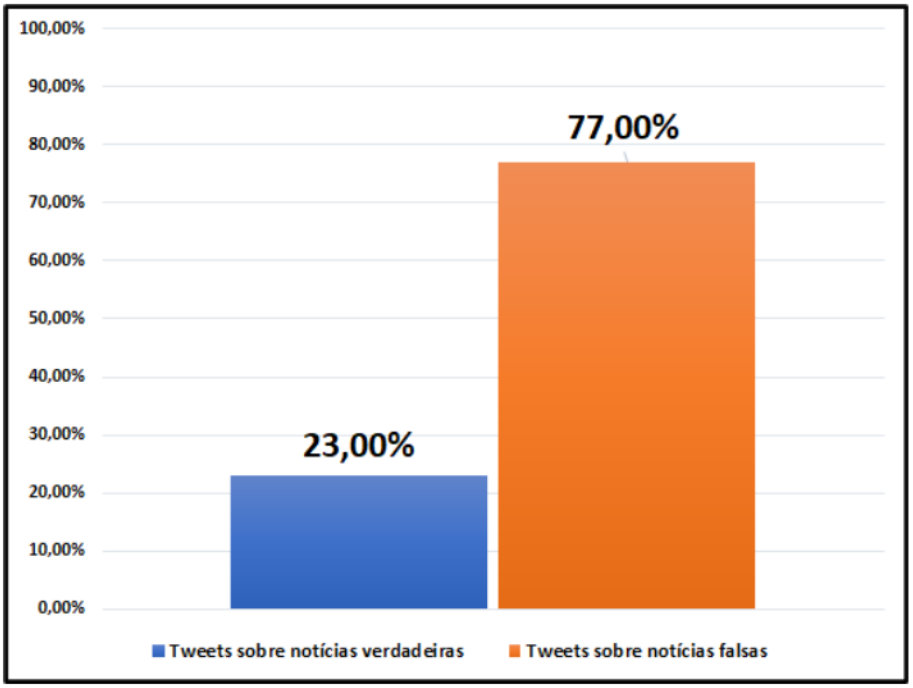
Figura 8 - Distribuição de tweets por rótulo dos seguidores de Fernando Haddad (PT)



Fonte: Elaboração do autor

Ao analisar a Figura 9, é possível observar que os seguidores do candidato Jair Bolsonaro também publicaram mais *tweets* sobre *fake news*, e em uma proporção ainda maior que a dos seguidores do candidato do PT.

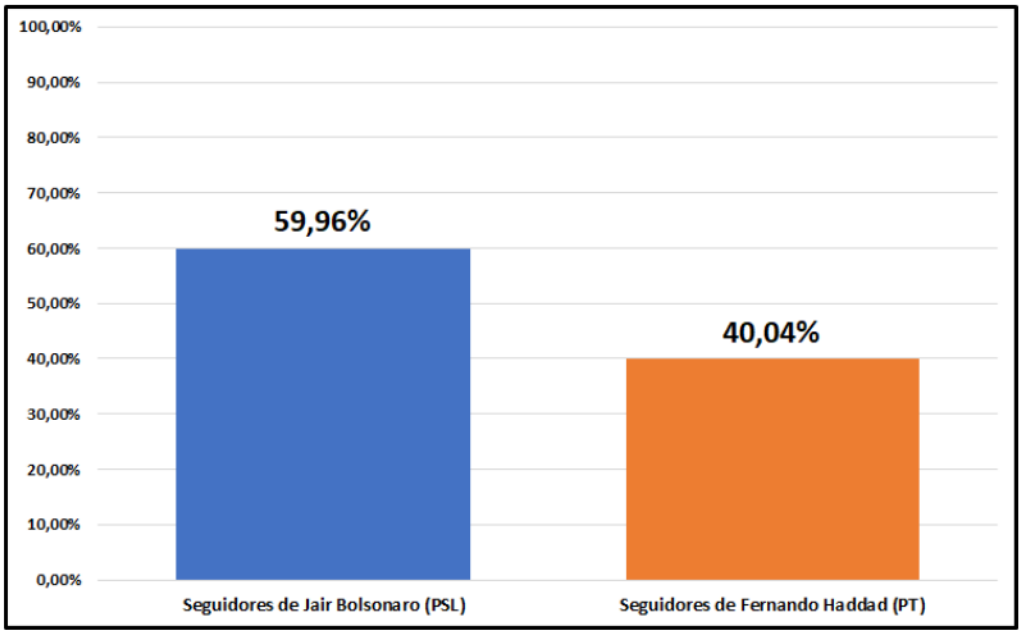
Figura 9 - Distribuição de tweets por rótulo dos seguidores de Jair Bolsonaro (PSL)



Fonte: Elaboração do autor

Por fim, também consideramos uma outra dimensão de análise do perfil dos seguidores. Para contextualizar essa nova análise, é importante salientar que, inicialmente, foram coletados os seguidores e, depois, foram obtidos seus *tweets*, gerando um espaço de tempo entre as duas coletas. Ao realizar a coleta dos *tweets*, um fato que reforça a tese de que as redes sociais têm sido uma ferramenta para a disseminação de conteúdo falso chamou a atenção. Dos 2.310.218 perfis coletados, 485.098 (20,99%) não existiam mais, no momento da obtenção dos *tweets*. Tal fato evidencia uma característica dos chamados *robôs*, utilizados para propagar *fake news*, ou seja, um perfil é criado, usado para divulgar notícias e rapidamente é apagado da rede social. Na Figura 10, é possível observar a distribuição das contas excluídas em um curto espaço de tempo, por seguidores de cada candidato.

Figura 10 - Porcentagem de contas excluídas rapidamente por seguidores de cada candidato



Fonte: Elaboração do autor

7.2 Ameaças à Validade

Embora os resultados do experimento tenham se mostrado satisfatórios, os mesmos apresentam ameaças à sua validade que devem ser comentadas.

Ameaças à validade externa:

Os resultados demonstrados, na Figura 7, são baseados em notícias da base de dados disponível e no período analisado. Esta base, por sua vez, contém informações das três principais agências de *fact-checking* do Brasil, no entanto, outras *fake news* sobre o contexto

das eleições presidenciais brasileiras de 2018 podem não ter sido checadas pelas agências e podem ter sido propagadas, ou seja, os percentuais se referem ao universo de notícias disponível e ao período analisado.

As informações coletadas do *Twitter* são de seguidores dos candidatos na referida rede social, todavia, não é possível afirmar que estes sejam de fato apoiadores e/ou eleitores destes candidatos. Também vale destacar que o percentual maior de *tweets* relacionados às *fake news* dos seguidores de um dos candidatos pode evidenciar um maior engajamento por parte destes nas redes sociais. Em outras palavras, não é possível evidenciar a intenção destes seguidores em divulgar *fake news*, uma vez que a prática de repassar uma notícia sem antes checá-la é muito comum.

Por fim, não existe nenhuma evidência encontrada por este experimento de que os candidatos tenham apoiado ou incentivado a proliferação dessas informações falsas.

Ameaças à validade de construção:

As implementações dos métodos comparados por este estudo devem atender aos requisitos teóricos, assim, alterações podem comprometer seus resultados. Desta forma, visando garantir implementações corretas, utilizou-se a biblioteca *Scikit-Learn* (PEDREGOSA; VAROQUAUX *et al.*, 2011), a qual possui citações em estudos relacionados.

8. Conclusão

Ao replicar um experimento, este trabalho contribuiu para a consolidação da base de conhecimento existente sobre o processo de detecção de *fake news*. Atualmente, a disseminação de *fake news* é um problema presente nos mais diversos contextos e, ao seguir um processo experimental, este trabalho colaborou com futuras replicações em outros contextos. Isto é, uma base de conhecimento robusta só poderá ser gerada com as replicações de verdadeiros experimentos controlados que validem estatisticamente seus trabalhos, as quais poderão servir de insumo para verdadeiras metanálises dos dados.

Neste contexto, uma das principais dificuldades na realização deste tipo de experimento é a obtenção dos dados. Desta forma, a fim de contribuir com a comunidade e dar transparência a esta pesquisa, foram disponibilizados, no (KAGGLE, 2018), três conjuntos de dados: o primeiro, contendo as notícias já rotuladas sobre as eleições presidenciais brasileiras de 2018, o segundo, (2) contendo todos os *tweets* coletados para o experimento, e o terceiro, (3) com os

tweets manualmente e arduamente classificados, utilizados para averiguar a visão geral de *fake news* por seguidores.

Para essa última base de dados extraída, transformada e carregada, os resultados mostraram que existem diferenças significativas entre os métodos utilizados, e que, apesar do TF-IDF possuir a maior média de eficácia, verificou-se que, estatisticamente, este possui similaridade com o BM25. Portanto, respectivamente para o TF-IDF e BM25, os resultados alcançados para classificação de *fake news* foram: Acurácia (79,86% e 79,00%), Precisão (79,97% e 78,76%), Sensibilidade (78,97% e 76,05%) e Medida-F1 (79,47% e 77,38%). Para os métodos de incorporação de palavras, os resultados poderão ser melhores ao haver uma combinação com técnicas de *Deep Learning*, bem como pode ser a indicação da necessidade de novas pesquisas e adaptações destes métodos para o contexto das *fake news*.

Ao comparar os resultados desta pesquisa com o trabalho publicado em (JIN; CAO *et al.*, 2017), notou-se que, em ambos os trabalhos, os métodos TF-IDF e BM25 tiveram resultados semelhantes. No presente trabalho, a ausência de significância estatística foi evidenciada ao utilizar o Teste de *Tukey*. Além disso, considerando o universo de notícias checadas disponível, o período analisado e uma margem de erro de 3,5%, também se evidenciou que ambos os lados políticos divulgaram *fake news* e que seguidores do candidato Jair Bolsonaro (PSL) foram responsáveis por 62,25% dos *tweets* relacionados a *fake news*, contra 37,75% dos seguidores do candidato Fernando Haddad (PT). No que diz respeito às contas excluídas da rede social em um curto espaço de tempo, 59,96% eram de seguidores do candidato do PSL e 40,04% de seguidores do candidato do PT. Vale ressaltar que a divulgação de *fake News* não implica necessariamente intenção, isto pode estar relacionado apenas a um maior engajamento de seguidores de um candidato.

Por fim, como trabalhos futuros, pretende-se expandir a análise para outros modelos. Além disso, o desenvolvimento de uma aplicação web capaz de receber um texto como entrada e devolver a notícia mais próxima, bem como sua classificação no que diz respeito à veracidade.

REFERÊNCIAS

AL-ANZI, Fawaz S.; ABUZEINA, Dia. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. **Journal of King Saud University-Computer and Information Sciences**, v. 29, n. 2, p. 189-195, 2017. ALLCOTT, Hunt; GENTZKOW, Matthew. Social media and fake news in the 2016 election. Journal of economic perspectives, p. 211-36, 2017.

AMAZON. **Amazon Web Services**, 15 out. 2019. Disponível em: <https://aws.amazon.com/pt/>. Acesso em 15 de Outubro de 2019

ANJOS, A. **Análise de Variância**, 2009, Acessado em 18 de Outubro de 2019. Disponível em: <http://www.est.ufpr.br/ce003/material/apostilace003.pdf>.

BASILI, Victor R.; WEISS, David M. A methodology for collecting valid software engineering data. **IEEE Transactions on software engineering**, n. 6, p. 728-738, 1984.

BIRD, Steven. **NLTK**, 01 de set .de 2020. Disponível em: <https://www.nltk.org/>. Acesso em : 01 de Setembro de 2020.

BUCKLEY, Chris. The importance of proper weighting methods. In: **Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993**.

CAELEN, Olivier. A Bayesian interpretation of the confusion matrix. **Annals of Mathematics and Artificial Intelligence**, v. 81, n. 3, p. 429-450, 2017.

CASTILLO, Carlos; MENDOZA, Marcelo; POBLETE, Barbara. Information credibility on twitter. In: **Proceedings of the 20th international conference on World wide web**. 2011. p. 675-684..

CIAMPAGLIA, Giovanni Luca et al. Computational fact checking from knowledge networks. **PloS one**, v. 10, n. 6, p. e0128193, 2015.

COLLINS. Collins Dictionary. **Collins**, 25 mar. 2017. Disponível em: <https://www.collinsdictionary.com/word-lovers-blog/new/collins-2017-word-of-the-year-shortlist,396,HCB.html>. Acesso em: 25 de Março de 2017.

CONROY, Nadia K.; RUBIN, Victoria L.; CHEN, Yimin. Automatic deception detection: Methods for finding fake news. **Proceedings of the association for information science and technology**, v. 52, n. 1, p. 1-4, 2015.

CONTRATRES, Felipe. Similaridade entre títulos de produtos com Word2Vec. **Medium**, 2020. Disponível em: <https://medium.com/luizalabs/similaridade-entre-t%C3%ADtulos-de-produtos-com-word2vec-5e26199862f0>. Acesso em: 16 Novembro de 2020.

DAVID, Lazer et al. The science of fake news. **Science**, v. 359, n. 6380, p. 1094-1096, 2018..

DEKKER, Gerben W.; PECHENIZKIY, Mykola; VLEESHOUWERS, Jan M. Predicting Students Drop Out: A Case Study. **International Working Group on Educational Data Mining**, 2009.

ELASTIC. Practical BM25 - Part 2: The BM25 Algorithm and its Variables, 2020. **Elasitc**. Disponível em: <https://www.elastic.co/pt/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables>. Acesso em: 15 Outubro de 2020.

FATOS, Aos. **Aos Fatos**, 14 ago. 2018. Disponível em: <https://www.aosfatos.org/>. Acesso em: 15 de Agosto de 2018.

FERNANDES, Hugo Miguel Moutinho. As novas guerras: o desafio da guerra híbrida. **Revista de Ciências Militares**, v.4, 2016.

FIELD, Andy. **Descobrindo a estatística usando o SPSS-5**. Penso Editora, 2009.

FRIEDMAN, Milton. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the american statistical association**, v. 32, n. 200, p. 675-701, 1937.

FRIGGERI, Adrien et al. **Rumor Cascades**. Eighth International AAAI Conference on Weblogs and Social Media. Michigan: [s.n.]. 2014. p. 101-110.

GUPTA, Manish; ZHAO, Peixiang; HAN, Jiawei. **Evaluating event credibility on twitter**. Proceedings of the 2012 SIAM International Conference on Data Mining. [S.l.]: SIAM. 2012. p. 153-164.

HAN, Jiawei; PEI, Jian; TONG, Hanghang. **Data mining: concepts and techniques**. Morgan kaufmann, 2022.

HAND, David J. Principles of data mining. **Drug safety**, v.30, n.7, p.621-622, 2007.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. **Springer**, 2011.

JIN, Zhiwei et al. News credibility evaluation on microblog with a hierarchical propagation model. In: **2014 IEEE international conference on data mining**. IEEE, 2014. p. 230-239.

JIN, Zhiwei et al. **Detection and analysis of 2016 us presidential election related rumors on twitter**. International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. [S.l.]: Springer. 2017. p. 14-24.

KAGGLE. **Kaggle**, 18 out. 2018. Disponível em: < https://www.kaggle.com/caiovms/datasets?scroll=true>. Acesso em: 18 de Outubro de 2018.

LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. In: **International conference on machine learning**. PMLR, 2014. p. 1188-1196.

LEVENE, H. Robust tests for equality of variances. **International Journal of Machine Learning and Cybernetics**, 1960. 278-292.

LI, Baoli; HAN, Liping. Distance weighted cosine similarity measure for text classification. **International Conference on Intelligent Data Engineering and Automated Learning**. Springer. 2013. p. 611-618.

LUHN, Hans P. The automatic creation of literature abstracts. **IBM Journal of research and development**, p. 159-165, 1958.

LUPA, Agência. **Agência Lupa**, 20 out. 2019. Disponível em: <https://piaui.folha.uol.com.br/lupa/>. Acesso em: 20 de Outubro de 2019.

MACHADO, Emerson Lopes. Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes. 2007..

MAKICE, Kevin. **Twitter API: Up and running: Learn how to build applications with the Twitter API**. " O'Reilly Media, Inc.", 2009.

MÁRQUEZ-VERA, Carlos; MORALES, Cristóbal R.; SOTO, Sebastian V. Predicting school failure and dropout by using data mining techniques. **IEEE Revista Iberoamericana de Tecnologias del Aprendizaje**, p. 7-14, 2013.

MIKOLOV, Tomas et al. Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, v. 26, 2013.

MONGODB. **MongoDB**, 15 jan. 2020. Disponível em: <https://www.mongodb.com/>. Acesso em: 15 de Janeiro de 2020.

MORRIS, Meredith Ringel et al. Tweeting is believing? Understanding microblog credibility perceptions. In: **Proceedings of the ACM 2012 conference on computer supported cooperative work**. 2012. p. 441-450.

MUNDSTOCK, Elsa et al. Introdução à Análise Estatística utilizando o SPSS 13.0. **Cadernos de Matemática e Estatística Série B. Universidade Federal do Rio Grande do Sul, Porto Alegre, RS**, 2006.

REIS, Julio CS; BENEVENUTO, Fabrício. Supervised Learning for Misinformation Detection in WhatsApp. In: **Proceedings of the Brazilian Symposium on Multimedia and the Web**. 2021. p. 245-252.

OGAWA, Taro. **Num2Words**, 09 jan. 2020. Disponível em: <https://pypi.org/project/num2words/>. Acesso em: 09 de Janeiro de 2018.

OLIVEIRA, Robert A N de; COLAÇO JÚNIOR, Methanias. Experimental analysis of stemming on jurisprudential documents retrieval. **Information**, 2018. 28.

ÖZCAN, Said. **Tweet Pre-Processor**, 01 set. 2020. Disponível em: <https://pypi.org/project/tweet-preprocessor/>. Acesso em: 01 de Setembro de 2018.

PATRO, V M.; PATRA, Manas R. Augmenting weighted average with confusion matrix to enhance classification accuracy. **Transactions on Machine Learning and Artificial Intelligence**, p.77-91, 2014.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. **The Journal of machine Learning research**, p. 2825-2830, 2011.

PÚBLICA, Agência. Agência Pública, 08 maio 2018. Disponível em: <https://apublica.org/>.

ROBERTSON, Stephen; ZARAGOZA, Hugo. The probabilistic relevance framework: BM25 and beyond. **Foundations and Trends in Information Retrieval**, p. 333-389, 2009.

ROCHLIN, Nick. Fake news: belief in post-truth. **Library hi tech**, p. 386-392, 2017.

RONG, Xin. Word2Vec parameter learning explained. **ArXiv preprint arXiv:1411.2738**, 2014.

RUEDIGER, Marco A. et al. Robôs, redes sociais e política no Brasil: estudo sobre interferências ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018. **FGV DAPP**, 2017.

SALTON, Gerard; WONG, Anita; CHUNG-SHU, Yang. A vector space model for automatic indexing. **Communications of the ACM**, p. 613-620, 1975.

SALTON, Gerrard; BUCKLEY, Christopher. Term-weighting approaches in automatic text retrieval. **Information processing & management**, p. 513-523, 1988.

SANTOS, Rafael Meneses et al. **Long Term-short Memory Neural Networks and Word2vec for Self-admitted Technical Debt Detection**. ICEIS. [S.l.]: [s.n.]. 2020. p. 157-165.

SEWARD, Lori E.; DOANE, David P. **Estatística Aplicada à Administração e Economia-4**. AMGH editora, 2014.

SHAPIRO, S.S; WILK, M.B. An Analysis of Variance Test for Normality (Complete Samples). **International Journal of Machine Learning and Cybernetics**, 1965. 591-611.

SPINELLI, Egle M.; ALMEIDA SANTOS, Jéssica. Jornalismo na era da pós-verdade: fact-checking como ferramenta de combate às fake news. **Revista Observatório**, p. 759-782, 2018.

SPSS. **IBM SPSS software**, 25 out. 2020. Disponível em: <https://www.ibm.com/analytics/spss-statistics-software>. Acesso em: 25 de Outubro 2020.

STATISTA. **Statisa,** Most popular social networks worldwide as of January 2022, ranked by number of monthly active users, 20 mar. 2019. Disponível em: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Acesso em: 12 de Outubro 2019.

TIAN, Yuan; LO, David; SUN, Chengnian. **Information retrieval based nearest neighbor classification for fine-grained bug severity prediction**. 2012 19th Working Conference on Reverse Engineering. [S.l.]: IEEE. 2012. p. 215-224.

TRAVASSOS, Guilherme Horta; GUROV, Dmytro; AMARAL, E. A. G. G. **Introdução à engenharia de software experimental**. UFRJ, 2002.

TUMASJAN, Andranik et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. In: **Proceedings of the International AAAI Conference on Web and Social Media**. 2010. p. 178-185..

TWITTER. Twitter muda regras para combater fake news e manipulação política, 20 mar. 2019. Disponível em: <https://help.twitter.com/pt/rules-and-policies/twitter-report-violation>. Acesso em: 20 de Março de 2020.

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The spread of true and false news online. **Science**, p. 1146-1151, 2018.

WANG, Hao et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: **Proceedings of the ACL 2012 system demonstrations**. 2012. p. 115-120.

WHATSAPP. O WhatsApp continua pessoal e privado, 07 set. 2020. Disponível em: <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>. Acesso em: 07 de Setembro de 2020.

WILCOXON, Frank. Individual comparisons by ranking methods. In: **Breakthroughs in statistics**. Springer, New York, NY, 1992. p. 196-202..

WOHLIN, Claes et al. **Experimentation in software engineering**. Springer Science & Business Media, 2012.

WU, Ke; YANG, Song; ZHU, Kenny Q. False rumors detection on sina weibo by propagation structures. In: **2015 IEEE 31st international conference on data engineering**. IEEE, 2015. p. 651-662.

ZHAO, Zhe; RESNICK, Paul; MEI, Qiaozhu. Enquiring minds: Early detection of rumors in social media from enquiry posts. In: **Proceedings of the 24th international conference on world wide web**. 2015. p. 1395-1405.

Original recebido em: 19 de novembro de 2020

Aceito para publicação em: 21 de julho de 2022

Caio Vinícius Meneses Silva

Graduado em Sistemas de Informação pela Universidade Federal de Sergipe. Possui Mestrado em Ciência da Computação pelo Programa de Pós-Graduação em Ciência da Computação (PROCC) também pela Universidade Federal de Sergipe. Além disso, atua como Engenheiro de Software na iniciativa privada a cerca de 10 anos.