UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Análise Exploratória e Experimental sobre Detecção Inteligente de Fake News

Dissertação de Mestrado

Caio Vinícius Meneses Silva

São Cristóvão – Sergipe

2020

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Caio Vinícius Meneses Silva

Análise Exploratória e Experimental sobre Detecção Inteligente de Fake News

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Prof. Dr. Methanias Colaço Rodrigues Júnior

São Cristóvão – Sergipe

2020

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Silva, Caio Vinícius Meneses
S586a Análise exploratória e experimental sobre detecção inteligente de fake news / Caio Vinícius Meneses Silva ; orientador Methanias Colaço Rodrigues Júnior. – São Cristóvão, SE, 2020.
80 f.

Dissertação (mestrado em Ciências da Computação) – Universidade Federal de Sergipe, 2020.

1. Fake news. 2. Eleições. 3. Processamento eletrônicos de dados. I. Rodrigues Júnior, Methanias Colaço, orient. II. Título.

CDU 004.421.4:070.16

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: CAIO VINÍCIUS MENESES SILVA

Em 08 dias do mês de dezembro do ano de dois mil e vinte, com início às 10h00min, realizou-se na Sala virtual https://meet.google.com/buc-qjii-wtu. A Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Caio Vinicius Meneses Silva**, que desenvolveu o trabalho intitulado: **"Análise Exploratória e Experimental sobre Detecção Inteligente de Fake News"**, sob a orientação do Prof. Dr. **Methanias Colaço Rodrigues Júnior**. A Sessão foi presidida pelo Prof. Dr. **Methanias Colaço Rodrigues Júnior** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Profª. Drª. **Claudia Cappelli Aló** (UFRJ), logo em seguida o Prof Dr **Rogério Patrício Chagas do Nascimento** (PROCC/UFS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) aprovado "(aprovado/reprovado)". Atendidas as exigências da Instrução Normativa 01/2017/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), Resolução nº 25/2014/CONEPE e da Portaria nº 413 de 27 de maio de 2020 (Banca por videoconferência) que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária *"Prof. José Aloísio de Campos"*, 08 de dezembro de 2020.

Methanias C. R. Júnior
Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

Prof. Dr. Rogério Patrício Chagas do Nascimento
(PROCC/UFS)
Examinador Interno

Claudia Cappelli
Profª. Drª. Claudia Cappelli Aló
(UFRJ)
Examinador Externo

Caio Vinícius Meneses Silva
Caio Vinicius Meneses Silva
Candidato

Dedico este trabalho a todos os que
acreditam que através da educação, podemos construir um país melhor.

Agradecimentos

Primeiramente, agradeço a Deus por me dar saúde, força e sabedoria para que pudesse concretizar mais essa etapa em minha vida.

Em seguida, agradeço a toda minha família, que sempre me apoiou e me incentivou nos estudos, especialmente meus pais, que sempre me deram as condições necessárias para tal.

A minha noiva Tamires, que além de suportar minha ausência devido as atividades acadêmicas, sempre foi minha incentivadora, muito obrigado!

Um agradecimento especial ao meu professor e orientador Methanias por ser um exemplo e estar sempre presente, por seus sábios ensinamentos e por me fazer admirar ainda mais a profissão, muito obrigado mesmo!

Agradeço ainda ao cara que me despertou o interesse em ingressar no mestrado, ainda durante a graduação, Dr. Gilton Mal, obrigado!

Ao meu colega de Mestrado Rafael Silva Fontes e meu ex-colega de trabalho e amigo Felipe Machado (Harry) por todo auxílio durante o trabalho, muito obrigado!

Por fim, a todos que indiretamente ou diretamente, estiveram envolvidos nesse processo e contribuíram de alguma forma para que chegasse até aqui, o meu mais sincero agradecimento a todos vocês.

Uma mentira pode dar a volta ao mundo,
enquanto a verdade ainda calça seus sapatos.
(Mark Twain)

Resumo

Contexto: A evolução dos meios de comunicação tem contribuído para a disseminação de notícias falsas, principalmente após o surgimento das redes sociais digitais. No entanto, esta prática não é um fenômeno recente na história da humanidade. Relatos do período da Primeira Guerra Mundial evidenciam o uso de propaganda enganosa por parte da imprensa, que culminou em novas normas de objetividade e equilíbrio jornalístico. Nas mídias sociais digitais, tal fenômeno, agora chamado de *fake news*, encontrou um novo ambiente propício para se espalhar em escalas mundiais, tornando 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 *fake news*. **Objetivo:** Este trabalho teve por propósito avaliar a eficácia dos métodos mais utilizados para verificar correspondência de textos, na tarefa de detecção automática de *fake news* sobre as eleições presidenciais brasileiras de 2018, comparando as evidências encontradas com os resultados obtidos de um mapeamento do estado da arte publicado nesta pesquisa. **Método:** Inicialmente, foi realizado um mapeamento sistemático para identificar e caracterizar as principais abordagens, técnicas e algoritmos usados, na computação, para a detecção de notícias falsas. Por fim, foi realizado um experimento controlado, *in vitro*, usando como perspectiva um dos trabalhos encontrados na literatura, cujo contexto possui forte relação com este estudo: as eleições americanas de 2016. Desta forma, avaliou-se a eficácia dos métodos, confrontando os resultados e os contextos dos dois trabalhos. **Resultados:** Para o estado da arte, foi identificado que os principais algoritmos utilizados na tarefa de detecção de notícias falsas são LSTM (17,14%), Naive-Bayes e Algoritmo de Similaridade (11,43% cada um). Com a execução de todo o processo experimental, foi evidenciado que os métodos TF-IDF e BM25 obtiveram médias estatisticamente similares de acurácia, respectivamente, 79,86% e 79,00%. Por fim, os métodos *Word2Vec* e *Doc2Vec* obtiveram resultados um pouco abaixo dos demais, 75,69% e 72,39% respectivamente. **Conclusões:** Após a análise do estado da arte, evidenciou-se lacunas relacionadas a trabalhos no contexto *Big Data* e à necessidade de replicações dos estudos existentes, na forma de experimentos mais controlados. Com a avaliação experimental, foi constatado que as eficácias dos métodos avaliados foram similares às eficácias do trabalho utilizado como controle. Além disso, considerando o universo de notícias checadas disponível, o período analisado e uma margem de erro de aproximadamente 3,5%, evidenciou-se a divulgação de *fake news* da parte de seguidores de ambos os candidatos avaliados no experimento. Os seguidores do candidato Jair Bolsonaro (PSL) foram responsáveis por 62,25% dos *tweets* relacionados a notícias falsas, 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, em alguns casos indica apenas um maior engajamento.

Palavras-chave: Eleições, Mineração de Texto, Experimentação e *Fake News*.

Abstract

Context: The evolution of the media has contributed to the spread of false news, especially after the emergence of digital social networks. However, this practice is not a recent phenomenon in human history. Reports from the First World War period show the use of misleading advertising by the press, which culminated in new standards of objectivity and journalistic balance. In digital social media, this phenomenon, now called fake news, has found a new environment conducive to spreading worldwide, making it impossible to manually check this immense volume 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 fake news. **Objective:** The purpose of this work was to evaluate the effectiveness of the most used methods to check text correspondence, in the task of automatic detection of fake news about the Brazilian presidential elections of 2018, comparing the evidence found with the results obtained from a mapping of the state of art published in this research. **Method:** Initially, a systematic mapping was carried out to identify and characterize the main approaches, techniques and algorithms used, in computing, to detect false news. Finally, a controlled experiment was carried out, in vitro, using as perspective one of the works found in the literature, whose context has a strong relationship with this study: the American elections of 2016. In this way, the effectiveness of the methods was evaluated, comparing the results and contexts of the two works. **Results:** For the state of the art, it was identified that the main algorithms used in the task of detecting false news are LSTM (17.14%), Naive-Bayes and Similarity Algorithm (11.43% each). With the execution of the entire experimental process, it was evidenced that the TF-IDF and BM25 methods obtained statistically similar averages of accuracy, respectively, 79.86% and 79.00%. Finally, the Word2Vec and Doc2Vec methods also obtained, respectively, the worst averages, 75.69% and 72.39%. **Conclusions:** After analyzing the state of the art, gaps related to work in the Big Data context and the need for replication of existing studies, in the form of more controlled experiments, became evident. With the experimental evaluation, it was found that the effectiveness of the methods evaluated were similar to the effectiveness of the work used as a control. In addition, considering the universe of checked news available, the analyzed period and a margin of error of approximately 3.5%, the disclosure of fake news by the followers of both candidates evaluated in the experiment was evidenced. Followers of 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 period of 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.

Keywords: Elections, Text Mining, Experimentation and Fake News.

Sumário

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 usuários 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; RUBIN; CHEN, 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 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 notícias falsas 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 (LAZER 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, 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 usuário 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, 2018). 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; 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 avaliar a eficácia dos métodos mais utilizados para verificar correspondência de textos, na tarefa de detecção automática de *fake news* eleitorais, comparando as evidências encontradas com os resultados obtidos de um mapeamento do estado da arte, cujo objetivo paralelo foi identificar e caracterizar as principais abordagens, técnicas e algoritmos usados,

na computação, para a detecção de notícias falsas. Para a avaliação dos métodos supracitada, foi realizado um experimento controlado, *in vitro*, usando como perspectiva um dos trabalhos encontrados na literatura, cujo contexto possui forte relação com este estudo.

Em se tratando dos resultados gerais, para o estado da arte, foi identificado que os principais algoritmos utilizados na tarefa de detecção de notícias falsas são LSTM (17,14%), *Naive-Bayes* e Algoritmo de Similaridade (11,43% cada um). Com a execução de todo o processo experimental, foi evidenciado que os métodos TF-IDF e BM25 obtiveram médias estatisticamente similares de acurácia, respectivamente, 79,86% e 79,00%. Por fim, os métodos *Word2Vec* e *Doc2Vec* obtiveram resultados um pouco abaixo dos demais, também respectivamente, 75,69% e 72,39%.

Na próxima seção, será introduzida a problemática e hipóteses relacionadas à pesquisa em questão.

1.1 Problemática e Hipótese

O massivo uso de mídias sociais, pelas mais diversas classes sociais e níveis de escolaridade, tem contribuído com a divulgação das *fake news* nos ambientes virtuais (DAVIS; PROCTOR, 2017). Diversos meios de comunicação alertam sobre o perigo das notícias falsas, uma vez que buscam confundir fatos, no intuito de prejudicar a compreensão correta por parte da sociedade.

No campo político, a situação é ainda mais complexa, uma vez que este tipo de desinformação ou contrainformação é utilizada para favorecer um determinado candidato ou partido, no contexto das eleições. Ainda não se sabe de fato até onde podem chegar as consequências da divulgação de *fake news*, mas algumas teorias já atribuem a estas a interferência direta no resultado de eventos políticos recentes (RECUERO; GRUZD, 2019).

Diante desse contexto e visando conhecer o estado da arte acerca das abordagens de Computação Inteligente utilizadas para a detecção de *fake news* nas mídias sociais, foi realizado um Mapeamento Sistemático da Literatura (KITCHENHAM; CHARTERS, 2007). A pesquisa nas bases digitais retornou um total de 153 artigos e, ao final da análise, restaram 35 para a extração dos resultados. De acordo com o mapeamento, além dos resultados resumidos na seção anterior, a mapeamento permitiu averiguar que, direta ou indiretamente, por se tratar de processamento de textos, os trabalhos sempre lidam com algum método de mapeamento do texto em um vetor numérico. Neste sentido, esses métodos foram um ponto comum e proeminente para análise, instigando a realização da avaliação da qualidade dos mais utilizados.

Um outro resultado encontrado no mapeamento realizado foi a ausência de experimentos completos. Apesar de muitos estudos terem elaborado bons projetos experimentais e boas estruturas para validação dos dados, nenhum validou a significância dos resultados. Em outras

palavras, do ponto de vista científico, não podemos classificá-los como experimentos controlados. Estes resultados reforçam as evidências de que esta área carece de experimentos controlados que provejam pacotes experimentais capazes de serem replicados, aumentando assim a base de conhecimento sobre o assunto. Os resultados completos do mapeamento podem ser vistos no Capítulo 2.

Dentre os trabalhos analisados no mapeamento, notou-se que o estudo descrito em (JIN et al., 2017) tem uma relação direta com esta dissertação: as eleições norte-americanas de 2016, uma vez que este trabalho baseou-se nas eleições brasileiras de 2018, e a avaliação de métodos de mapeamento de textos, também um dos interesses desta pesquisa. Do ponto de vista das eleições, ambos os pleitos foram marcados pela polarização do eleitorado, por uma intensa onda de disseminação de *fake news* nas mídias sociais, além da suspeita de interferência de outros países na disputa pelo cargo de presidente.

Detalhando um pouco mais o trabalho relacionado supracitado (JIN et al., 2017), os autores propõem a classificação 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 (verdadeira ou falsa). 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 mesmo estudo, 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.

Considerando a similaridade entre os contextos desta dissertação e do trabalho supracitado, o qual não dispõe de uma metodologia de análise experimental completa, realizou-se um experimento controlado, dentro do contexto das eleições brasileiras. A realização de experimentos similares em diferentes contextos é uma característica importante para qualquer tipo de estudo em computação, uma vez que a base de conhecimento específica se fortalece e ganha maior significância. O processo experimental proporciona, de modo sistemático, disciplinado e controlado, a avaliação de processos e de atividades humanas (TRAVASSOS; GUROV; AMARAL, 2002). A experimentação ajuda a determinar a eficácia de métodos e de teorias propostas. Somente experimentos verificam as teorias existentes, podem explorar os fatores críticos, e dar luz a um novo fenômeno, para que as novas teorias possam ser formuladas (BASILI; SHULL; LANUBILE, 1999).

Neste contexto, para guiar o estudo, foi elaborada a seguinte questão principal de pesquisa, cuja resposta visa cumprir um dos objetivos deste trabalho:

No contexto da detecção de notícias falsas eleitorais no *Twitter*, entre os métodos de mapeamento utilizados para correspondência de texto selecionados no Mapeamento Sistemático

da Literatura, qual o melhor em termos das métricas de qualidade a serem avaliadas (Acurácia, Precisão, Sensibilidade e Medida-F1)?

Sendo assim, com o objetivo principal e métricas definidas, será considerada a hipótese a seguir (para cada métrica):

- H_0: Os métodos possuem médias iguais para a métrica.
  $\mu 1$(metrica) = $\mu 2$(métrica) ... = μn(métrica);
- H_1: Os métodos possuem médias diferentes para a métrica.
  $\mu 1$(metrica) $\neq$ $\mu 2$(métrica) ... $\neq$ μn(métrica);

1.2 Justificativa

Uma *fake news* tem 70% mais chances de ser compartilhada que uma notícia verdadeira (VOSOUGHI; ROY; ARAL, 2018). Desta forma, informações falsas ganham espaço na internet de forma mais rápida, mais profunda e com mais abrangência que as verdadeiras. Notícias falsas tendem a ser mais impactantes ou inéditas, o que acaba atraindo pessoas, que movidas pela sensação de privilégio ou ineditismo, acabam divulgando a informação.

Este tipo de conteúdo enganoso desencadeia consequências complexas para a sociedade, podendo gerar reações não previstas e até tragédias. Um dos casos mais conhecidos ocorreu nos EUA, durante as eleições presidenciais de 2016, e ficou conhecido como *"Pizzagate"*. Sites criados por apoiadores do então candidato Donald Trump espalharam boatos de que sua concorrente, a senadora e candidata à presidência, Hillary Clinton, seria líder de uma rede de prostituição e tráfico infantil, e que os abusos aconteciam no porão de uma pizzaria chamada *Comet Ping Pong*, localizada em Washington. O boato, que começou em fóruns e sites, e migrou rapidamente para redes sociais tais como *Facebook* e *Twitter*, espalhou-se de tal forma que dele resultaram investigações conduzidas pela polícia local e por renomados jornais, bem como investigações virtuais, feitas por cidadãos indignados com o suposto crime. Um desses cidadãos decidiu investigar pessoalmente a rede de exploração sexual, levando consigo três armas e efetuando três disparos, que, felizmente, não atingiram nenhuma família presente no local (TEIXEIRA et al., 2018).

Outro exemplo, também relacionado com a política, é caso *Brexit*, termo formado pela junção das palavras *Britain* e *exit*, usado para se referir ao referendo que decidiu sobre a saída do Reino Unido da União Europeia. A votação foi marcada pela confusão dos eleitores, supostamente causada pelas *fake news*. Mesmo com sérias dificuldades para realizar o acordo, as consequências oriundas da saída influem diretamente na modificação de leis e estruturas governamentais daquele país, influenciando empresas, organizações e a vida dos cidadãos (GILCHRIST, 2018). Desta forma, considerando a natureza duvidosa, a dificuldade em julgá-las como verdadeiras ou falsas, o volume e velocidade com que essas informações são disseminadas

pelas redes sociais, bem como seus impactos na sociedade, evidencia-se a necessidade da criação de ferramentas que venham auxiliar na determinação da autenticidade dessas informações de forma automática.

1.3 Objetivos

Para realização desta pesquisa, têm-se os seguintes objetivos geral e específicos.

1.3.1 Objetivo Geral

O objetivo deste trabalho é caracterizar o estado da arte da área de detecção inteligente de *fake news* e avaliar a eficácia dos métodos mais utilizados para verificar correspondência de textos, na tarefa de detecção automática de *fake news* sobre as eleições presidenciais brasileiras de 2018, comparando as evidências encontradas com os resultados obtidos em um mapeamento sistemático da literatura.

1.3.2 Objetivos Específicos

Para possibilitar a realização do objetivo geral, listamos os seguintes objetivos específicos:

- Mapeamento Sistemático da Literatura, com a finalidade de identificar os principais algoritmos, métodos e técnicas de classificação encontrados nos trabalhos sobre detecção de notícias falsas;
- Experimento controlado para avaliação de métodos utilizados para correspondência de textos, no contexto das discussões sobre as eleições presidenciais brasileiras de 2018 encontradas no *Twitter*.

1.4 Metodologia

Este trabalho é um estudo experimental que avaliou a eficácia dos métodos mais utilizados para verificar correspondência de textos, na tarefa de detecção automática de *fake news* sobre as eleições presidenciais brasileiras de 2018. A fim de avaliar tais questões, foram utilizadas as seguintes métricas: (i) Acurácia, (ii) Precisão, (iii) Sensibilidade e (iv) Medida-F1.

Além do ponto de vista experimental, este trabalho também se caracteriza como exploratório, uma vez que, inicialmente, foi realizado um Mapeamento Sistemático da Literatura, publicado em (SILVA; FONTES; JÚNIOR, 2020), com o objetivo de identificar e sistematizar as principais abordagens, técnicas e algoritmos usados, na computação, para a detecção de notícias falsas. A análise desses estudos demonstrou a ausência da aplicação de uma metodologia rigorosa e experimental nas avaliações publicadas, bem como de trabalhos no contexto de *Big*

Data. Visando contribuir com a experimentação, nosso estudo foi consolidado em diversas fases: planejamento, instrumentação, seleção de participantes, preparação do ambiente, execução, coleta de dados e validação estatística dos resultados.

Para a execução do experimento, foram coletadas informações publicadas no *Twitter* de seguidores dos dois principais candidatos a presidente, nas eleições brasileiras de 2018, além de notícias sobre estas eleições, previamente checadas em sites de *fact-checking*. Em seguida, foi analisado o esquema de correspondência de texto usado em (JIN et al., 2017), no qual é possível medir a similaridade entre dois documentos, ou entre um documento específico e todo o *corpus*. O design experimental pode ser visualizado na seção 5 do Capítulo 3. Por conseguinte, para auxiliar os cálculos e verificar se havia diferenças significativas na eficácia dos tratamentos, foram utilizados seis testes estatísticos: *Anova, Friedman, Levene, Shapiro-Wilk, Tukey* e *Wilcoxon*.

Em resumo, este trabalho conduziu um estudo exploratório e um experimento, os quais têm os seus métodos descritos de forma autocontida, nos seus planejamentos, detalhados nos capítulos 2 e 3.

1.5 Organização da Dissertação

Este documento está organizado de acordo com a Instrução Normativa Nº 02/2015/PROCC, a qual permite que a Dissertação seja "uma compilação de artigos científicos submetidos ou publicados em veículos com *Qualis*, desde que seja contextualizada com seções de Introdução e Conclusão, não limitada a estas". São 4 capítulos que fornecem uma base conceitual e experimental para o entendimento sistêmico. Os tópicos a seguir descrevem o conteúdo de cada um dos capítulos:

- O Capítulo 1 apresenta esta Introdução, explicando as justificativas e as hipóteses levantadas;
- O Capítulo 2 traz um artigo que descreve o mapeamento sistemático publicado no *Journal of Applied Security Research*, Qualis A1 para área de Direito;
- O Capítulo 3 traz um artigo que disserta sobre a avaliação experimental, submetido à Revista Interamericana de Comunicação Midiática - Animus, Qualis A3;
- Finalmente, no capítulo 4, é apresentado um compilado de conclusões, contribuições e sugestões de trabalhos futuros.

2

Mapeamento Sistemático

Este capítulo traz um artigo do mapeamento sistemático, publicado no *Journal of Applied Security Research*, Qualis A1.

JOURNAL OF APPLIED SECURITY RESEARCH
https://doi.org/10.1080/19361610.2020.1761224

Intelligent Fake News Detection: A Systematic Mapping

Caio V. Meneses Silva, Raphael Silva Fontes, and Methanias Colaço Júnior

Computing Department, Federal University of Sergipe, São Cristóvão, Brazil

ABSTRACT

Context: The speed with which the Fake News spread today has encouraged work in various areas to minimize the damage and the public insecurity caused by their proliferation.
Objective: To characterize and analyze Fake News threat detection.
Method: Systematic Mapping, since the area youthfulness still prevents a complete meta-analysis.
Results: The most used algorithms were LSTM (17.14%), Naive-Bayes and Similarity Algorithm (11.43%).
Conclusions: There is still the absence of more controlled experiments in the Big Data context. Fake News is a national security problem, requiring effective solutions to combat it. Situations like the Covid-19 virus (coronavirus) reinforce this fact.

KEYWORDS
Computational and Artificial Intelligence; machine learning; fake news; public security

1. Introduction

In the age of smartphones, more and more people are using social networks and platforms such as Facebook, Whatsapp, and Twitter, together, have about 4 billion users worldwide (Statista, 2019). Checking notifications, sending, and receiving content have become a daily task, changing the way news are published and consumed. This phenomenon and all these interactions performed by users around the world generate a huge mass of data called "Big Data" (Conroy et al., 2015).

One consequence of this new way of accessing information is the increased range and volume of this data, and, despite the advantages provided by these mechanisms, the spread of false news has presented itself as a current problem, because, given the amount and speed with which they spread, checking the truth of the facts has become a difficult task (Ciampaglia et al., 2015).

The High-tech data analytics, coupled with ultra-sophisticated personality testing based on social networking activity have been used to produce and direct Fake News to highly specific sections of the population to influence

CONTACT Caio V. Meneses Silva kaiovinicius_@hotmail.com Computing Department, Federal University of Sergipe, Marechal Rondon Avenue, s/n, Rose Elze Garden, São Cristóvão, Sergipe 49100-000, Brazil.
This article has been republished with minor changes. These changes do not impact the academic content of the article.

people in the most diverse segments, such as politics and public security (Allcott & Gentzkow, 2017; Belova & Georgieva, 2018). Since 2016, the use of the term "Fake News" has increased by 365%, making it the word of the year 2017, by Collins, the traditional English language dictionary (Collins, 2017). This fact has drawn the attention of researchers and large companies that have collaborated in the search for ways to combat this practice (Tandoc et al., 2018).

Facebook was the first social network that intended to implement a solution to reduce the reach of Fake News. In order to do this, a new option has been created where you can flag an item in your news feed as fake, so, this way, if enough users rate a news item as fake, besides appearing less in other user's news feed, whoever posted the fake content will receive a warning, informing that many people have warned that the news is unreliable and contains misleading information (Rochlin, 2017).

Twitter also announced in 2018 a stricter set of rules to curb the spread of Fake News and political manipulation on the social network (Twitter, 2018). Due to its potential for circulating news content, microblogging has been used as a strategic part of divulgating false information, and its use as a tool for journalism is increasingly gaining ground in the online environment, especially as it is a fast and interactive channel for the divulgation of real-time content in a short way.

Other tools, such as fact checking, or even websites, have also gained strength lately. Fact checking stories with data, surveys and records, it is also a way of qualifying the public debate through the journalism, besides checking the degree of veracity of the information (Myslinski, 2012).

The Artificial Intelligence (A.I.) has been a paradox when it comes to Fake News. While intelligent robots contribute to the automatic dissemination of this kind of information, Machine Learning algorithms have contributed to their automatic detection (Ruediger et al., 2017). Machine Learning (M.L.) is an area of Artificial Intelligence that uses algorithms that can make pattern predictions, given a data set, using patterns that have been discovered to predict future data or to make a decision. The data used for learning are defined by numerical variables extracted from relevant characteristics (Murphy, 2012).

The dissemination of false content is not something unheard of in the history of mankind. Some events are reported throughout history, such as the use of propaganda by journalists in World War I, which culminated in new norms of objectivity and journalistic equilibrium (Lazer et al., 2018). However, it was in digital social media that such phenomenon, now called Fake News, found a new enabling environment to spread in global scales, causing way much losses to society.

The dubious nature of news disseminated by social networks, combined with the difficulty of judging it as true or false, fosters a need for the creation of tools to assist in determining the authenticity of such information, automatically. In this context, this paper presents a Systematic Mapping (S.M.) whose objective was to identify and characterize the algorithms used to detect false news, in the Big Data context or not, using articles from important databases.

After answering the research questions, it was found that the most used algorithms were: Long Short-Term Memory, with 17.14%, Naive-Bayes and Similarity Algorithm, with 11.43%. In the context of empirical evaluations, the "Case Study," with 69% of publications, far exceeded the other types of validations analyzed. In relation to countries, the United States, China and India lead the ranking of publications on the subject. The analysis of the publications in relation to the years in which they were published shows that from 2016 there was a substantial increase in research on the subject, 32 publications were found during this period, which corresponds to 92% of the articles. Among the publicity vehicles, conferences stood out, with 71% of publications.

The rest of this article is organized as it follows. Section 2 reports the absence of related work. In Section 3, the method adopted in this paper is addressed. Section 4 presents and discusses the results achieved. Section 5 summarizes implications and lessons learned. In Section 6, the threats to validity encountered are detailed. And finally, in Section 7, the conclusion is presented.

2. Related Works

No studies were found to map the research on technologies applied to combat Fake News. This highlights a gap and the need to compile all the related works, identifying the current state of the area and what may be done.

3. Method

Systematic Mapping (S.M.) is a study that aims to identify, evaluate and interpret all available research relevant to a particular research question (Kitchenham & Charters, 2007). It is a rigorous study based on systematic literature review methods that aim to structure the area under investigation (Petersen et al., 2008). The steps we performed in this mapping study are detailed in Figure 1.

Figure 1. The steps of a systematic mapping process.

Thus, following the protocol proposed by Kitchenham and Charters (2007), the objective of this study is to perform an MS aiming to characterize scientific works on the use of intelligent Fake News detection algorithms, techniques, and approaches. The definition of the research questions, their scope, the search strategy and the selection criteria will be described in the following sections.

3.1. Research Questions

The research questions were developed with the purpose of presenting an overview of the area, highlighting key aspects of the primary studies. For this study, such questions attempt to provide a specific insight into the relevant aspects of detecting Fake News in the Big Data context or in any application and social media, regardless of the volume, variety or speed of information production. In addition to including questions about which algorithms are most commonly used for Fake News detection, it was also asked which countries and years that most publish about the topic, the main publication vehicles and which empirical evaluations were made (controlled experiment, case study, proof of concept and practical application).

Controlled Experiment is a form of experimental study in which the researcher has control over the main aspects of the study and the independent variables being studied. Its objective is to confirm theories, conventional knowledge, to evaluate the prediction of models or to validate measurements. It involves the formulation of hypotheses to be verified in relation to the results obtained (Wohlin et al., 2012).

According to Yin (2001), the Case Study is a research strategy that encompasses a method that encompasses everything in specific approaches to data collection and analysis. It is useful when the phenomenon to be studied is broad and complex and cannot be studied outside the context in which it occurs naturally. In addition, it investigates a contemporary phenomenon, starting from its real context, using multiple sources of evidence.

Proof of concept is a term used to refer to a practical model that can prove the (theoretical) concept established by a research or technical article. It can also be considered a generally summarized or incomplete

Table 1. PICO model for compliance of research questions.

Category	Description
Population	Publications by researchers and developers for Fake News analysis
Intervention	Applications using smart approaches, techniques and algorithms for detecting Fake News in the Big Data context.
Control	Applications that detect Fake News without the use of intelligence. Search String base controls articles. Papers from applications that obey the intervention: * Rumor Gauge: Predicting the Veracity of Rumors on Twitter (Vosoughi et al., 2017) * Evaluating Machine Learning Algorithms for Fake News Detection (Gilda, 2018) * In Search of Credible News (Hardalov et al., 2016) Papers from the most common applications that do not fit the intervention: * Fake News and The Economy of Emotions (Bakir & McStay, 2018) * Fake News and Journalism Education (Richardson, 2017) * Fake news, post-truth and media–political change (Corner, 2017)
Results	Fake news prediction and detection in an automated manner.

Table 2. Research questions.

No	Description
RQ1	What algorithms are most used in detecting Fake News in the Big Data context?
RQ2	Witch empirical evaluations are used?
RQ3	Which countries have the most researchers publishing on this topic?
RQ4	In what years were more works published in this area?
RQ5	What are the main popular media?

implementation of a method or idea, made with the purpose of verifying that the concept or theory in question is susceptible to be exploited in a useful way (de Freitas Farias et al., 2017).

Finally, for the classification made in this paper, practical applications consist of implementations without scientific or experimental validations. An even simpler implementation than a proof of concept, with no real minimum evaluation.

As a guarantee of research questions quality, the PICO (Population, Intervention, Control and Outcomes) model was used (James et al., 2016), whose objective is to prove the ability to characterization and classification of the issues. With this model, it is also possible to evaluate the effects of an intervention on a given population. A control environment with pre-mapped articles has been defined to validate the outcome of the questions and the search terms as shown in Table 1.

In Table 2, it is possible to observe the research questions that were defined for this work., Questions 3 and 4 characterize the research in the area. Questions 2 and 5, in addition to characterizing, allow assessing the maturity of the current research stage, showing if there is the need to increase the use of scientific method in this area, with replications of studies that will allow to evaluating if other researchers independently will come up with the same results. Question 1 will help researchers to select algorithms for their applications. Finally, due to the hypothesis of the small

Table 3. Categories of the PICO model and terms.

Category	Description
Population	Fake News
Intervention	Big Data, Data Mining, Text Mining, Smart Computing, Intelligent Computing, Artifical Intelligence, Natural Language Processing, NLP, Machine Learning
Control	Investigative journal analysis on Fake News (no strings)
Results	Approach, Technique, Method, Algorithm, Detection, Classification, Prediction, Discovery, Application, System, Tool, Framework

number of replications (new area), a question to ascertain the quality of the algorithms was not raised. Answering this type of question requires an analysis of the external validity of similar works and the combination of experimental evidence, considering contexts and biases. However, in the synthesis analysis section, the best current results are presented.

3.2. Search Scope

Five databases were used to perform the Systematic Mapping, they are ACM, Scopus, Engineering Village, IEEE, and Web of Science. The access to the bases was made through the Higher Education Personnel Improvement Coordination Portal (CAPES) with the signature of the educational institution, so it was possible to consult texts that have exclusive subscriber content.

Although Scopus encompasses works from other bases, it is not proven that its content contains all reference about the query, therefore, individual queries were held, aiming to contemplate a larger amount of works.

3.3. Publication Search Method

To perform the search in digital bases, a search string was defined using English terms and the use of various terms, associated with the assumption that the studies would be contained in the areas of computation that deal with Fake News detection. These terms were identified with the help of the PICO model control articles, described in Section 3.1, which were later refined and adapted to make the most of the string. Table 3 shows the terms, before refining them, that were selected.

After refinement, the adjusted terms were used to construct the search string, which are described in Table 4.

The search string generated with the terms highlighted above was:

(Approach OR Technique* OR Method* OR Algorithm* OR Detection* OR Classification* OR Prediction* OR Discovery OR Application* OR System* OR Tool* OR Framework) AND ("Big Data" OR" Data Mining" OR" Text*

Table 4 Terms used in the search string.

Search strings terms		
Approach*		
Technique*		
Method*	Big Data	
Algorithm*	Data Mining	
Detection*	Text Mining	
Classification*	Smart Computing	
Prediction*	Intelligent Computing	Fake News
Discovery	Artificial Intelligence	
Application*	Natural Language Processing	
System*	NLP	
Tool*	Machine Learning	
Framework		

Mining" OR" Smart Computing" OR" Intelligent Computing" OR" Artifical Intelligence" OR" Natural Language Processing" OR NLP OR" Machine Learning") AND ("Fake News")

3.4. Selection Criteria

In order to filter the relevant documents for this Systematic Mapping, the inclusion and exclusion criteria were established. After the searches performed, using as basis the search string mentioned in Section 3.3, the collected results were counted taking into consideration only the studies selected for evaluation.

In order to apply the selection criteria, the summary and introduction of each article were analyzed. After this stage, the selected articles were read, analyzed and sent to the results extraction stage.

The inclusion criteria used were:

- Contain the theme of this study in the title, abstract or keywords;
- Explore some intelligent approach to detecting false news;
- Present some implementation of false news detection algorithm;
- Be available for online query.

The exclusion criteria used were:

- Publications that do not meet any inclusion criteria;
- Secondary studies, as they deal with third party approaches;
- Preliminary publications;
- Publications that do not concern the field of computer science;
- Publications that are not in English or Portuguese.

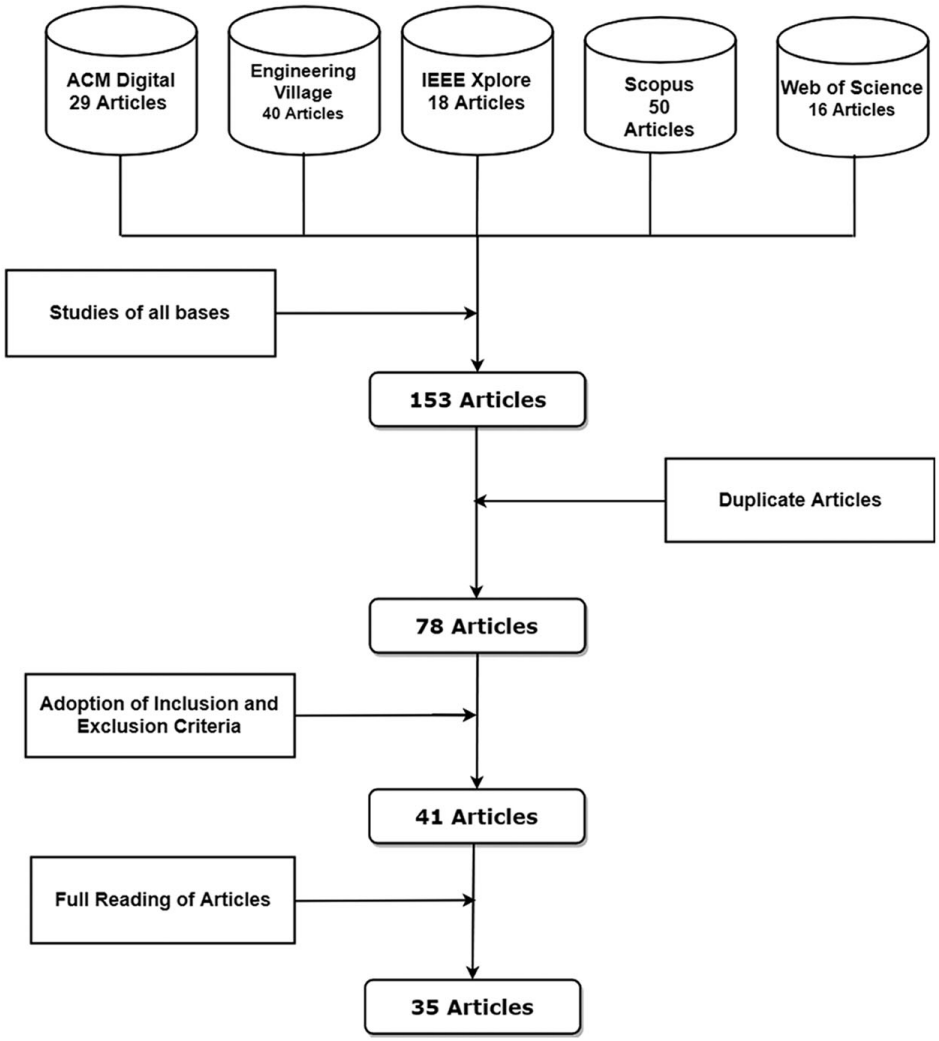

Figure 2. Selection process and job search.

4. Results and Discussions

In this section, the mapping process is demonstrated, from the search in the bases, to the selection of works for extraction and analysis of the results of the primary studies, until reaching in the answers to the research questions defined above. Figure 2 summarizes the mapping process.

Of the 153 articles, 78 were removed because they were duplicates. The inclusion and exclusion criteria were then applied and 34 articles were rejected, leaving a total of 41 papers for full reading and analysis. At this stage, it was found that 6 papers did not meet the requirements proposed by this mapping and were rejected according to the exclusion criteria. Figure 3 shows the quantity of articles by base after the selection step.

Figure 4 presents a graphic in response to research question **RQ1**. In this, it is possible to observe the most used algorithms in each individual assessment made by the articles accepted by this mapping. These are: Long Short-Term Memory (17.14%), Naive-Bayes and Similarity Algorithm

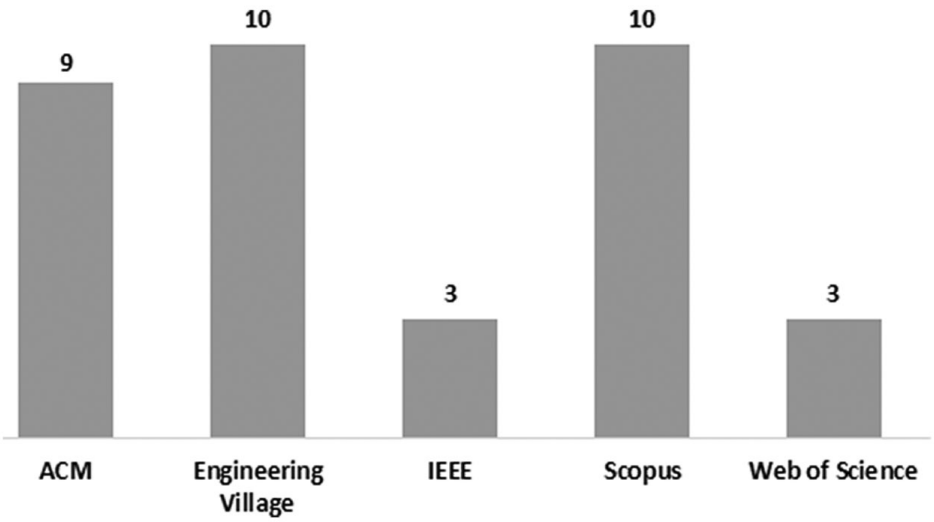

Figure 3. Distribution of primary studies by base.

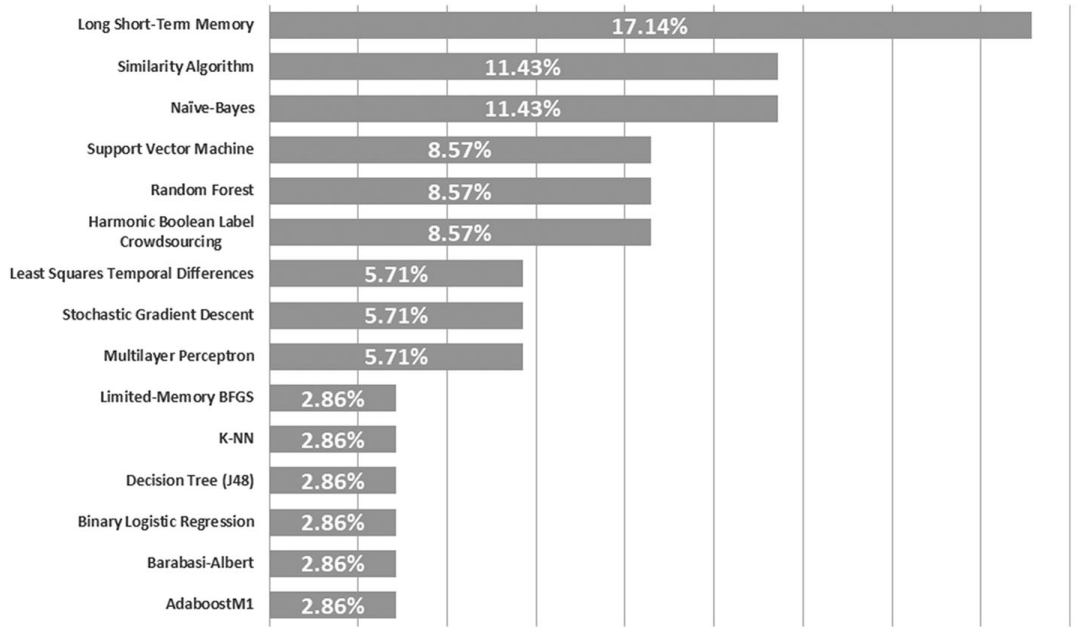

Figure 4. Distribution of primary studies by algorithms.

(11.43%), Support Vector Machine, Random Forest and Harmonic Boolean Label Crowdsourcing (8.57%), Least Squares Temporal Differences, Stochastic Gradient Descent and Multilayer Perceptron (5.71%), and LM-BFGS, K-NN, Decision Tree (J48), Binary Logistic Regression, Barabasi-Albert and AdaboostM1 (2.86%).

Table 5 summarizes the works and their respective references, as well as the algorithms that obtained the best results in each study.

Bhattacharjee et al. (2017) proposed a hybrid framework in which the classification task is divided between human supervision and a combination of algorithms and statistical methods. Thus, it was not possible to elect the best algorithm in this work.

Table 5. Articles, algorithms and references.

Article	Algorithm	References
Active Learning Based News Veracity Detection With Feature Weighting and Deep-Shallow Fusion	Proposed Method	(Bhattacharjee et al., 2017)
Adversarial Classification on Social Networks	Barabasi-Albert	(Yu et al., 2018)
An Algorithm for Supporting Decision Making in Stock Investment Through Opinion Mining and Machine Learning	Naïve-Bayes	(Jeong et al., 2018)
Arabic News Credibility on Twitter—An Enhanced Model Using Hybrid Features	Decision Tree (J48)	(Sabbeh & Baatwah, 2018)
Automatic Online Fake News Detection Combining Content and Social Signals	Harmonic Boolean Label Crowdsourcing	(Vedova et al., 2018)
Automatically Identifying Fake News in Popular Twitter Threads	Stochastic Gradient Descent	(Social networking (online), Buntain & Golbeck 2017)
Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers	Binary Logistic Regression	(Data mining, Aborisade & Anwar, 2018)
Combining Neural, Statistical and External Features for Fake News Stance Identification	Long Short-Term Memory	(Bhatt et al., 2018)
Contributions to The Study of Fake News in Portuguese—New Corpus and Automatic Detection Results	Multilayer Perceptron	(Monteiro et al., 2018)
CSI—A Hybrid Deep Model for Fake News Detection	Long Short-Term Memory	(Ruchansky et al., 2017)
Detect Rumor and Stance Jointly by Neural Multi-Task Learning	Random Forest	(Ma et al., 2018)
Detecting Journalistic Relevance on Social Media a Two-Case Study Using Automatic Surrogate Features	AdaboostM1	(Figueira & Guimarães, 2017)
Detection of Online Fake News Using N-Gram Analysis And Machine Learning Techniques	Support Vector Machine	(Ahmed et al., 2017)
Distrustrank—Spotting False News Domains	Similarity Algorithm	(Supervised learning, Woloszyn & Nejdl 2018)
Evaluating Machine Learning Algorithms for Fake News Detection	Stochastic Gradient Descent	(Gilda, 2018)
Fake News Detection—Network Data From Social Media Used To Predict Fakes	Harmonic Boolean Label Crowdsourcing	(Granskogen & Gulla, 2017)
Fake News Mitigation Via Point Process Based Intervention	Least Squares Temporal Differences	(Reinforcement learning, Farajtabar et al., 2017)
Filipino and English Clickbait Detection Using a Long Short Term Memory Recurrent Neural Network	Long Short-Term Memory	(Dimpas et al., 2018)
Generating Fake but Realistic Headlines Using Deep Neural Networks	Long Short-Term Memory	(Dandekar et al., 2017)
Identifying Tweets With Fake News	Support Vector Machine	(Krishnan & Chen, 2018)
In Search of Credible News	LM-BFGS	(Hardalov et al., 2016)

(*continued*)

Table 5. Continued.

Article	Algorithm	References
Inferring Trust From Message Features Using Linear Regression And Support Vector Machines	Naïve-Bayes	(Basharat & Ahmad, 2018)
Multimodal Fusion With Recurrent Neural Networks for Rumor Detection on Microblogs	Long Short-Term Memory	(Jin et al., 2017)
News Credibility Evaluation on Microblog With A Hierarchical Propagation Model	Support Vector Machine	(Jin et al., 2015)
News Reliability Evaluation Using Latent Semantic Analysis	Similarity Algorithm	(Xiaoning et al., 2018)
Organized Behavior Classification of Tweet Sets Using Supervised Learning Methods	Random Forest	(Beğenilmiş & Uskudarli, 2018)
Polarity Analysis of Editorial Articles Toward Fake News Detection	K-NN	(Samonte, 2018)
Ranking-Based Method for News Stance Detection	Multilayer Perceptron	(Zhang et al., 2018)
Rumor Gauge—Predicting the Veracity of Rumors On Twitter	Naïve-Bayes	(Vosoughi et al., 2017)
Rumor Detection on Twitter Pertaining to the 2016 U.S. Presidential Election	Similarity Algorithm	(Jin et al., 2017)
Searching for Diverse Perspectives in News Articles—Using an LSTM Network To Classify Sentiment	Long Short-Term Memory	(Harris, 2018)
Simple Open Stance Classification for Rumor Analysis	Random Forest	(Aker et al., 2017)
Some Like It Hoax—Automated Fake News Detection in Social Networks	Harmonic Boolean Label Crowdsourcing	(Tacchini et al., 2017)
Study of Hoax News Detection Using Naive Bayes Classifier In Indonesian Language	Naïve-Bayes	(Pratiwi et al., 2017)
The Diffusion of Misinformation on Social Media—Temporal Pattern, Message, and Source	Similarity Algorithm	(Shin et al., 2018)

As can be seen in Figure 4, among the 35 accepted papers, only Jin et al. (2017), which corresponds to 3% of the works, quote in its study the quantity and data analysis technology that characterize the use of Big Data.

In their work, Jin et al. (2017) propose classifying Fake News as a text matching task. For this, an empirical study was performed using similarity algorithms based on TF-IDF (Term Frequency—Inverse Document Frequency), BM25 (Best Matching 25), Word2Vec and Doc2Vec scores. After comparing the performance of unsupervised matching algorithms in a labeled set, the BM25 method was selected for the detection of Fake News-related tweets in the US presidential election in 2016. About 8 billion tweets were collected from around 14,000 followers of the two candidates and 1,723 Fake News rumors used in the correspondence task. In the end, the method achieved an accuracy of up to 94.7% (Figure 5).

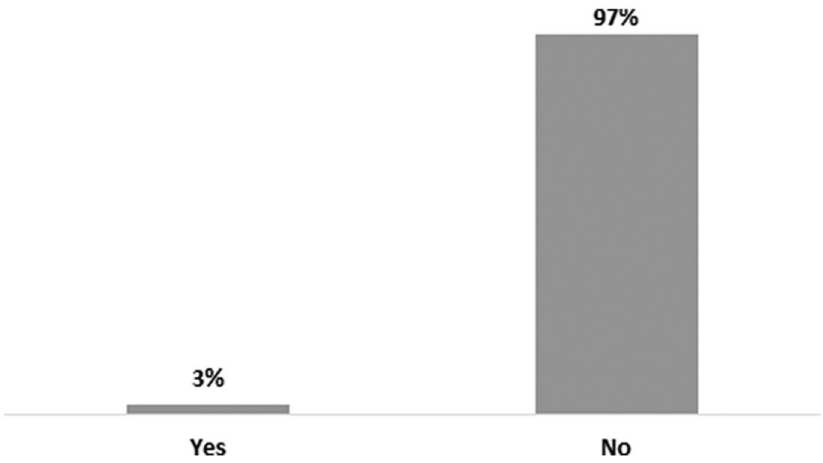

Figure 5. Distribution of primary studies using Big Data.

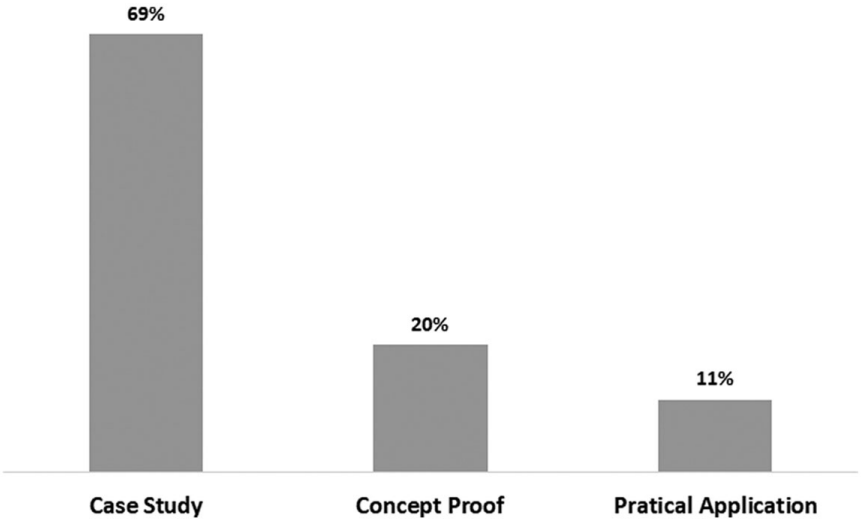

Figure 6. Distribution of primary studies by empirical evaluations.

In response to the **RQ2** research question, Figure 6 graphic shows the distribution of the primary studies according to the empirical evaluation they used. It is possible to observe that the Case Study is the main method used to validate the works, 69%. Proof of concept also appears in 20% of accepted papers. Finally, 11% of the articles used practical applications as a way to perform their work.

The absence of controlled experiments among the analyzed articles is a fact to be taken into consideration. This reflects a problem in the area of computer science, which is the absence of controlled experiments. Unlike other areas such as medicine, where a solution is exhaustively replicated, authors tend to propose solutions in computing, however, lack replication and controlled experimentation is common.

In response to the **RQ3** research question, as shown in Figure 7, we found that the United States was responsible for the largest number of publications, 10. Next, came China and India with three papers each. Italy,

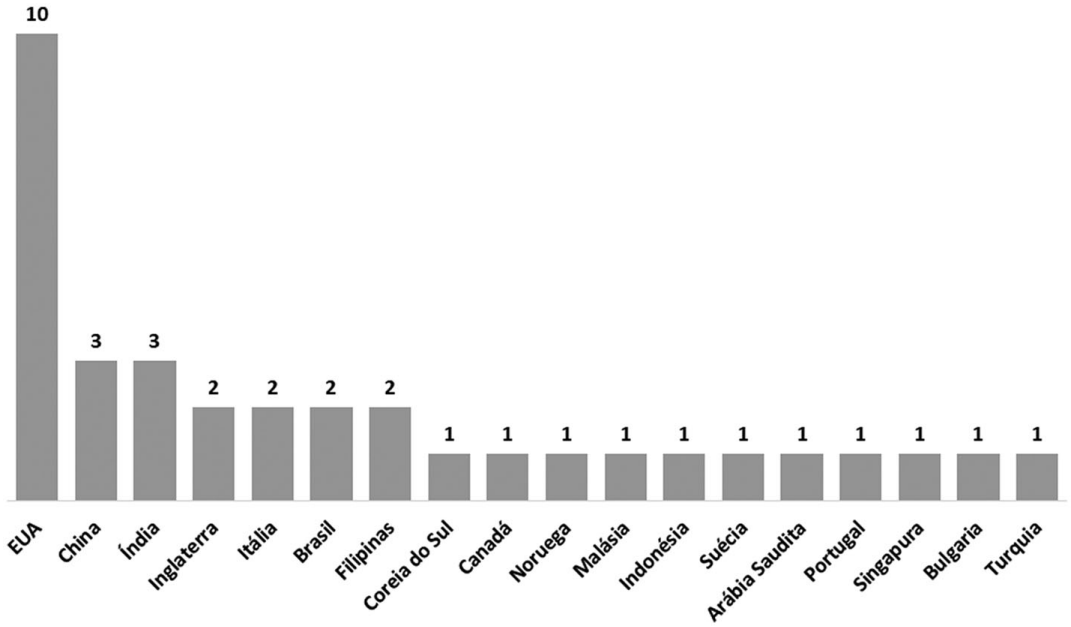

Figure 7. Distribution of primary studies by country.

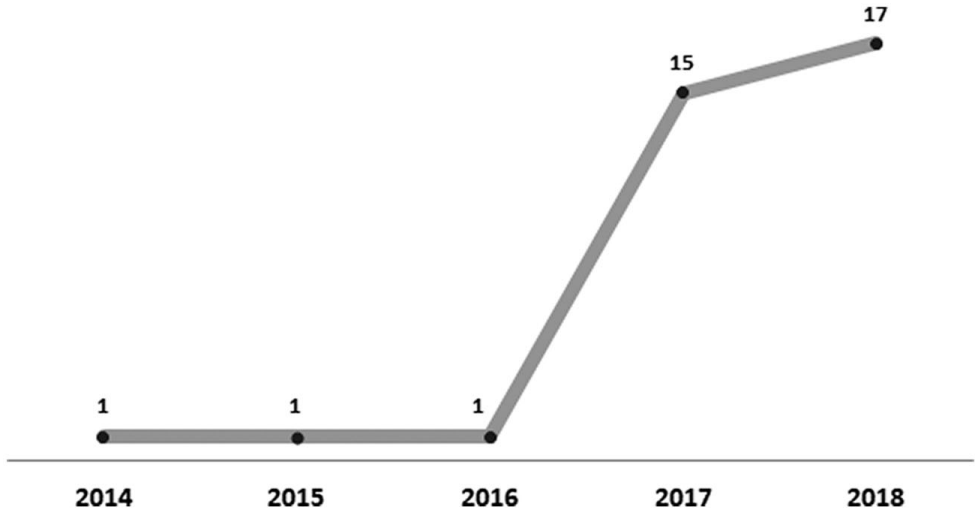

Figure 8. Distribution of primary studies per year.

England, Brazil and the Philippines appear on the list with two works each. And finally, South Korea, Canada, Norway, Malaysia, Indonesia, Sweden, Saudi Arabia, Portugal, Singapore, Bulgaria and Turkey complete the ranking with 1 article each.

In response to research question **RQ4**, Figure 8 shows the distribution of papers by year of publication.

From 2014 to 2016, 3 works were found, one each year. Since then, there has been a considerable increase in published research. In 2017, 15 works were mapped. In 2018 this growing continues, and 17 articles were found.

The cause of increased research on the topic may be related to the 2016 US presidential election. The election was marked by a huge wave of

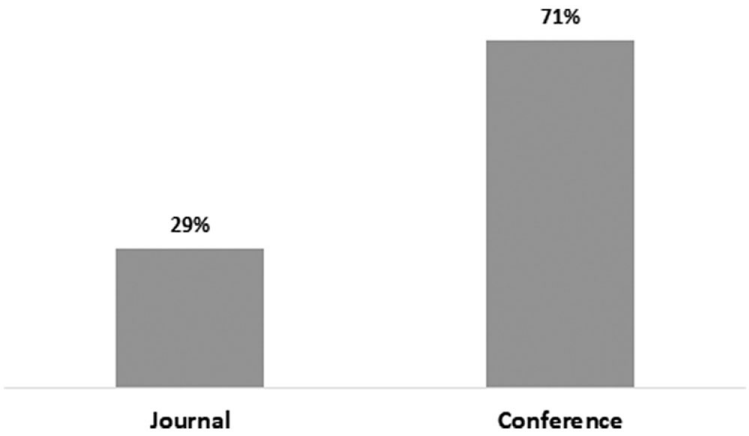

Figure 9. Distribution of primary studies by publication vehicle.

false news spreading on the Internet. It was also in this year that the term Fake News became popular, since then the academic community has turned its attention to this phenomenon, which now has social networks as an ally.

Looking at Figure 9, you can get the answer to **RQ5**. It is possible to notice that conferences were the most widely used medium of publication.

5. Syntheses Analysis

The low replication of studies still prevents a quantitative meta-analysis. However, from the point of view of the qualitative synthesis of the main aspects observed, it stands out the absence of evaluations in Big Data environments (de Oliveira Lima et al., 2019; Marr, 2015; McAfee et al., 2012), which have several essential characteristics, standing out the so-called 5 Vs, analyzed below from the Fake News point of view:

Volume: Refers to the large amount of information (brontobytes). The Big Data phenomenon background technology can handle this volume of data, storing and integrating it through specialized software and hardware. In addition to effectiveness, detecting Fake News also requires efficiency. In this sense, the most used algorithm (LSTM) requires processing power and needs to be evaluated with large volumes.

Velocity: It concerns the speed with which data and Fake News are created. Every moment, millions of messages on social networks or credit card transactions are checked. Big data technologies analyze data at the exact moment it is created, without the need to store it. This is a feature that needs to be combined with Fake News detection.

Variety: In the past, most of the data was structured and stored in tables and relations, today, the non-relational model is increasingly used to store messages, videos and sounds. With the Big Data world, this unstructured

data can be managed alongside traditional data. The question to be answered in the next studies will be the following: How to combine multimedia technologies, converters, audio transcriptions, image analysis and algorithms to deal with the variety of Fake News channels and to detect them?

Veracity: One of the most important characteristics of any information is the truth, especially in post-truth times. Initially, Big Data technologies did not allow the veracity control, however, combined with Artificial Intelligence, they can be a great alternative to solve this problem, allowing the construction of tools that automatically check the information authenticity.

Value: It is extremely important to access a massive amount of information every second, however, this becomes irrelevant, if it fails to generate value. Companies and governments need to join Big Data business, however, it is necessary to warn about the costs and benefits and try to add value to what is being done. If the Fake News application is not cost-effective, it should be discarded. In this context, research in this area still needs a more in-depth analysis of aggregated values and cost-benefits.

Despite the few replications of the studies, preventing deeper evaluations of bias and external validity, as well as the combination of experimental evidence, the value requirement guided the investigation and initial presentation of the algorithms that obtained the best accuracy performance, since the Big Data context requires solutions that consider the tradeoff effectiveness and efficiency. Below, we list the three best results, as well as the algorithm that achieved the worst performance.

The evaluation presented in Tacchini et al. (2017) achieved the best result, 99.40%, using the Harmonic Boolean Label Crowdsourcing algorithm. Then, Hardalov et al. (2016) obtained the second best result, 99.36%, using the LM-BFGS. The third best result was obtained by LSTM (Long Short-Term Memory), which was also the most used algorithm among all analyzed. With this algorithm, Ruchansky et al. (2017) achieved an accuracy of 95.30%. Finally, with the worst result, K-NN appears with an accuracy of only 40.00% (Samonte, 2018).

Regarding Public Security specific case, the spread of false content has an impact on several areas of society. Vasu et al. (2018), Haciyakupoglu et al. (2018), and Berghel (2017) treat Fake News as a national security problem. Vasu et al. (2018) and Haciyakupoglu et al. (2018), for example, refer to the use of Fake News as a means for organized disinformation campaigns, with the aim of destabilizing states through the subversion of societies. Vasu et al. (2018) still define this category as the most expensive, given its impact on countries' security and social cohesion. In many nations, this practice is already considered a crime, like Brazil, where the penalty for

those who disseminate false content for electoral purposes, for example, is two to eight years in prison (Ripoll & do Canto, 2019).

The results obtained in this work demonstrate that it is an area of interest for researchers worldwide and it has great potential. This study is relevant to the software companies and universities, fostering the need for interdisciplinary research between the areas of Computer Science and Security.

The research group that planned and executed this research will publish the experimental evaluation of the most effective algorithms mapped here, and an open tool, under development, which will be able to be coupled with a news portal and will help confirm the truth of the facts, preventing some criminals use methods such as modifying legitimate documents and distribute them as part of, for instance, disinformation campaigns.

6. Threats to Validity

Threats to validity may limit the ability to interpret and/or describe results from the data obtained. Therefore, there is no way to disregard the following threats found in this study.

6.1. Construction Validity

The elaborate search string may not completely cover the area of detection of Fake News in the Big Data context. To mitigate this threat, we sought to elaborate the search string as broadly as possible, using synonyms of terms used in the area. Following the PICO model, terms were identified and refined through control articles that were related to the search (intervention) and false positives, in order to calibrate the search string. In addition, the opinions of three researchers were considered.

6.2. Internal Validity

6.2.1. Data Extraction

Researchers were responsible for extracting and classifying the algorithms of each publication, bias or data extraction problem that could threaten the validity of the data characterization.

6.2.2. Selection Bias

Some articles may have been incorrectly categorized as articles were included or excluded in the systematic mapping according to the researchers' judgment. To mitigate these threats (Sections 6.2.1 and 6.2.2), selection

and extraction analyzes were made by the researchers involved, with a final vote on disagreements.

6.2.3. Classification Bias

Some selected articles did not make clear the methodology in detail, i.e., the research, evaluation or validation strategy. To mitigate this bias, these articles were read completely by the three researchers, in order to find characteristics that fitted a research type.

6.3. External Validity

Although the research was conducted in the main digital bases, it is impossible to say that the results of this systematic mapping covered all the works on the subject. However, this study presented evidence of the main techniques used and gaps to be explored, serving as a guide for future work in this line.

7. Conclusion

In this work, a mapping study was carried out to identify and analyze Intelligent Computing techniques used to detect false news in the Big Data context. The systematic mapping process was conducted using a study search and selection protocol that specified the method used in this work. Data were extracted and analyzed from 35 articles that met the chosen research line.

As a result, the most used algorithms for detecting Fake News were Long Short-Term Memory (17.14%), Naive-Bayes and Similarity Algorithm (11.43%), Support Vector Machine, Random Forest and Harmonic Boolean Label Crowdsourcing (8.57%), Least Squares Temporal Differences, Stochastic Gradient Descent and Multilayer Perceptron (5.71%), and LM-BFGS, K-NN, Decision Tree (J48), Binary Logistic Regression, Barabasi-Albert and AdaboostM1 (2.86%) (**RQ1**).

Most of the papers present empirical evidence, whether in case studies, used in 69% of the papers analyzed, and in proof of concept, but without rigorous validation. This can be explained by the fact that most publications are not yet done in journals. Consequently, the absence of controlled experiments as a form of empirical evaluation (**RQ2**) was also observed.

Among the primary studies selected for this mapping, we identified that the United States is the country that publishes the most on the topic. It was found that 10 out of 35 papers were published by the US, which corresponds to 29% of papers in this research line, China and India with three papers each following (**RQ3**).

Based on the analysis of the studies, it was observed that the existing research on the detection of false news had a considerable increase mainly from 2016. Among the studies analyzed, it was found that the year of 2018 was responsible for the largest number of articles, 17 out of 35 were published this year (**RQ4**).

When analyzing the selected papers in relation to the publications, it was observed that 71% of the studies were published in conferences and 29% in journals (**RQ5**).

Based on the results, it was observed that research on the subject is still very recent, so works in this line have a great potential for exploitation. A gap found in this mapping is the analysis of the theme in the context of voluminous data, since only one work cited the use of technologies and amount of data that characterized Big Data. A second shortcoming concerns the need for increased replication of studies already done, in the form of more controlled experiments, which will allow greater validation of results and new replications, more assertive, increasing the knowledge base on the subject and seeing the area of computing applied to public security as a Big Science. Fake news is a national security problem, requiring effective solutions to combat it. Situations like the Covid-19 virus—coronavirus—reinforce this fact.

ORCID

Caio V. Meneses Silva http://orcid.org/0000-0002-3242-660X
Raphael Silva Fontes http://orcid.org/0000-0003-3160-3384
Methanias Colaço Júnior http://orcid.org/0000-0002-4811-1477

References

Aborisade, O. M., Anwar, M. (2018). *Classification for authorship of tweets by comparing logistic regression and Naive Bayes classifiers*. 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 269–276). https://doi.org/10.1109/IRI.2018.00049

Ahmed, H., Traore, I., Saad, S. (2017). *Detection of online fake news using n-gram analysis and machine learning techniques*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10618 LNCS (pp. 127–138). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85032711154&doi=10.1007%2f978-3-319-69155-8_9&partnerID=40&md5=49365f16e11dc0f0be377b476e4169bc

Aker, A., Derczynski, L., Bontcheva, K. (2017). *Simple open stance classification for rumour analysis*. RANLP 2017 – Recent Advances in Natural Language Processing Meet Deep Learning (Vol. 2017-September, pp. 31–39). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045733701&doi=10.26615%2f978-954-452-049-6-005&partnerID=40&md5=00a3709d457181b781f30cc4aa1dc407

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. http://www.aeaweb.org/articles?id=10.1257/jep.31.2.211 https://doi.org/10.1257/jep.31.2.211

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Basharat, S., & Ahmad, M. (2018). Inferring trust from message features using linear regression and support vector machines. *Communications in Computer and Information Science*, *828*, 577–598. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049056246&doi=10.1007%2f978-981-10-8660-1_44&partnerID=40&md5=0fb953876e8e7c669605f651540c918

Beğenilmiş, E., Uskudarli, S. (2018). *Organized behavior classification of tweet sets using supervised learning methods*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053485153&doi=10.1145%2f3227609.3227665&partnerID=40&md5=774bfea7e02-f77278ee2c648f2fa67ca https://doi.org/10.1145/3227609.3227665

Belova, G., & Georgieva, G. (2018). Fake news as a threat to national security. *International Conference KNOWLEDGE-BASED Organization*, *24*(1), 19–22. https://doi.org/10.1515/kbo-2018-0002

Berghel, H. (2017). Alt-news and post-truths in the "fake news" era. *Computer Magazine.*, *50*(4), 110–114. https://doi.org/10.1109/MC.2017.104

Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., & Mittal, A. (2018). *Combining neural, statistical and external features for fake news stance identification*. Companion Proceedings of the Web Conference 2018 (pp. 1353–1357). https://doi.org/10.1145/3184558.3191577

Bhattacharjee, S. D., Talukder, A., & Balantrapu, B. V. (2017). *Active learning based news veracity detection with feature weighting and deep-shallow fusion*. 2017 IEEE International Conference on Big Data (Big Data), Boston, MA (pp. 556–565).

Buntain, C., Golbeck, J. (2017). *Automatically identifying fake news in popular Twitter threads*. 2017 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 208–215). https://doi.org/10.1109/SmartCloud.2017.40

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PLoS One*, *10*(6), e0128193. https://doi.org/10.1371/journal.pone.0128193

Collins. (2017). *Collins 2017 word of the year shortlist*. https://www.collinsdictionary.com/word-lovers-blog/new/collins-2017-word-of-the-year-shortlist,396,HCB.html

Conroy, N. J., Rubin, V. L., Chen, Y. (2015). *Automatic deception detection: Methods for finding fake news*. Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82).

Corner, J. (2017). *Fake news, post-truth and media–political change*. SAGE Publications Sage UK.

Dandekar, A., Zen, R., Bressan, S. (2017). Generating fake but realistic headlines using deep neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10439 LNCS* (pp. 427–440). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028462050&doi=10.1007%2f978-3-319-64471-4_34&partnerID=40&md5=e9ff204d06f547665bbb539ddb8a29fa

de Freitas Farias, M. A., Colaço Júnior, M., Spínola, R. O., & de Mendonça Neto, M. G. (2017). *Identifying technical debt through code comment analysis* (Unpublished doctoral dissertation). Universidade Estadual de Feira de Santana.

de Oliveira Lima, T., Júnior, M. C, de Jesus Prado, K. H., & dos S. Júnior, A. (2019). *A big data experiment to assess the effectiveness of deep learning neural networks in the mining of sustainable aspects of the hotels clients opinions*. 16th International Conference on Information Technology-New Generations (ITNG 2019) (pp. 201–207).

Dimpas, P., Po, R., Sabellano, M. (2018). *Filipino and English clickbait detection using a long short term memory recurrent neural network*. 2017 International Conference on Asian Language Processing (IALP), Singapore (Vol. 2018-January, pp. 276–280). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046716769&doi=10.1109%2fIALP.2017.8300597&partnerID=40&md5=72076a06871333c9f80099e7fa8df74a

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E, & Zha, H. (2017). Fake news mitigation via point process based intervention. *International Conference on Machine Learning*, *3*, 1823–1836.

Figueira, A., & Guimarães, N. (2017). *Detecting journalistic relevance on social media a two-case* study using automatic surrogate features. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 1136–1139). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040230568&doi=10.11452f3110025.3122120&partnerID=40&md5=1cddbd85c5104a23af6156f4-d6e4c703https://doi.org/10.1145/3110025.3122120

Gilda, S. (2018). *Evaluating machine learning algorithms for fake news detection*. 2017 IEEE 15th Student Conference on Research and Development (SCOReD) (Vol. 2018-January, pp. 110–115). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048881827&doi=10.1109%2fSCORED.2017.8305411&partnerID=40&md5=a37e929f8a87f1d7f20a8aa337f497e0

Granskogen, T., Gulla, J. (2017). *Fake news detection: Network data from social media used to predict fakes*. Proceedings of the 3rd Norwegian Big Data Symposium (NOBIDS 2017) (Vol. 2041, pp. 59–66). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041696530&partnerID=40&md5=6f55a3590cb4e65cabcc0236abe1f5d7

Haciyakupoglu, G., Hui, J. Y., Suguna, V., Leong, D., & Rahman, M. F. B. A. (2018). *Countering fake news: A survey of recent global initiatives*. S. Rajaratnam School of International Studies.

Hardalov, M., Koychev, I., & Nakov, P. (2016). *In search of credible news*. Artificial Intelligence: Methodology, Systems, and Applications: 17th International Conference, AIMSA 2016, Varna, Bulgaria, September 7–10, 2016, Proceedings (Vol. 9883, pp. 172–180).

Harris, C. (2018). *Searching for diverse perspectives in news articles: Using an LSTM network to classify sentiment*. ESIDA ’18, March 11, Tokyo, Japan (Vol. 2068). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044521756&partnerID=40&md5=a6baa9759145edc753dca83c357a4040

James, K. L., Randall, N. P., & Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. *Environmental Evidence*, *5*(1), 7. https://doi.org/10.1186/s13750-016-0059-6

Jeong, Y., Kim, S., & Yoon, B. (2018). *An algorithm for supporting decision making in stock investment through opinion mining and machine learning*. 2018 Portland International Conference on Management of Engineering and Technology (PICMET). (pp. 1–10). https://doi.org/10.23919/PICMET.2018.8481802

Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J. (2017). *Multimodal fusion with recurrent neural networks for rumor detection on microblogs*. Proceedings of the 2017 ACM on Multimedia Conference (pp. 795–816). http://doi.acm.org/10.1145/3123266.3123454 https://doi.org/10.1145/3123266.3123454

Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., & Luo, J. (2017). Rumor detection on twitter pertaining to the 2016 us presidential election. *arXiv preprint arXiv:1701.06250*.

Jin, Z., Cao, J., Jiang, Y.-G., Zhang, Y. (2015). *News credibility evaluation on microblog with a hierarchical propagation model*. 2014 IEEE International Conference on Data Mining, Shenzhen, China (Vol. 2015-January, pp. 230–239). https://www.scopus.com/inward/record.uri?eid=2-s2.0-84936948698&doi=10.1109%2fICDM.2014.91&partnerID=40&md5=7fca3f1de46b546f574383d6e5c00bd6

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Tech. Rep.). Technical report, EBSE Technical Report EBSE-2007-01.

Krishnan, S., & Chen, M. (2018). *Identifying tweets with fake news*. 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 460–464).

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science (New York, N.Y.)*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Ma, J., Gao, W., & Wong, K.-F. (2018). *Detect rumor and stance jointly by neural multi-task learning*. Companion Proceedings of the Web Conference 2018 (pp. 585–593). https://doi.org/10.1145/3184558.3188729

Marr, B. (2015). *Big data: Using smart big data, analytics and metrics to make better decisions and improve performance*. John Wiley & Sons.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60–68.

Monteiro, R., Santos, R., Pardo, T., Almeida, T. d., Ruiz, E., & Vale, O. (2018). *Contributions to the study of fake news in Portuguese: New corpus and automatic detection results*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11122 LNAI, 324–334. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053912872&doi=10.1007%2f978-3-319-99722-3_33&partnerID=40&md5=f4d89d0abb381443e2c5bb0ac96c3629

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Myslinski, L. J. (2012, May 22). *Fact checking method and system*. Google Patents. (US Patent 8,185,448)

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. *In Ease*, *8*, 68–77.

Pratiwi, I. Y. R., Asmara, R. A., & Rahutomo, F. (2017). *Study of hoax news detection using Naïve Bayes classifier in Indonesian language*. 2017 11th International Conference on Information Communication Technology and System (ICTS) (pp. 73–78).

Richardson, N. (2017). Fake news and journalism education. *Asia Pacific Media Educator*, *27*(1), 1–9. https://doi.org/10.1177/1326365X17702268

Ripoll, L., & do Canto, F. L. (2019). Fake news e" viralização": responsabilidade legal na disseminação de desinformação. *RBBD. Revista Brasileira de Biblioteconomia e Documentação*, *15*, 143–156.

Rochlin, N. (2017). Fake news: Belief in post-truth. *Library Hi Tech*, *35*(3), 386–392. https://doi.org/10.1108/LHT-03-2017-0062

Ruchansky, N., Seo, S., & Liu, Y. (2017). *CSI: A hybrid deep model for fake news detection*. Proceedings of the 2017 ACM on conference on information and knowledge management (pp. 797–806). http://doi.acm.org/10.1145/3132847.3132877

Ruediger, M. A., Grassi, A., Freitas, A., Contarato, A., Taboada, C., & Carvalho, D. (2017). others Robôs, redes sociais e política no brasil: estudo sobre interferências ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018.

Sabbeh, S., & Baatwah, S. (2018). Arabic news credibility on Twitter: An enhanced model using hybrid features. *Journal of Theoretical and Applied Information Technology*, *96*(8), 2327–2338. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046355216&partnerID=40&md5=43b9e4bb8ace4e8b1ae58e0e4b2d6cfa

Samonte, M. J. C. (2018). *Polarity analysis of editorial articles towards fake news detection*. Proceedings of the 2018 International Conference on Internet and e-Business (pp. 108–112). https://doi.org/10.1145/3230348.3230354

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, *83*, 278–287.

Statista. (2019). *Statista*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & Alfaro, L. D. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv: 1704.07506*.

Tandoc, E. C., Jr, Lim, Z. W., & Ling, R. (2018). Defining "fake news" a typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Twitter. (2018). *Twitter muda regras para combater fake news e manipulação política.* https://help.twitter.com/pt/rules-and-policies/twitter-report-violation

Vasu, N., Ang, B., Teo, T.-A., Jayakumar, S., Raizal, M., & Ahuja, J. (2018). *Fake news: National security in the post-truth era*. RSIS.

Vedova, M. L. D., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & Alfaro, L. d. (2018). *Automatic online fake news detection combining content and social signals*. 2018 22nd Conference of Open Innovations Association (FRUCT) (pp. 272–279). https://doi.org/10.23919/FRUCT.2018.8468301

Vosoughi, S., Mohsenvand, M. N., & Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on twitter. *ACM Transactions on Knowledge Discovery from Data*, *11*(4), 1–50: 36. http://doi.acm.org/10.1145/3070644 https://doi.org/10.1145/3070644

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Woloszyn, V., & Nejdl, W. (2018). *Distrustrank: Spotting false news domains*. The 10th ACM Conference, Amsterdam, Netherlands (pp. 221–228). http://dx.doi.org/10.1145/3201064.3201083

Xiaoning, G., De Zhern, T., King, S., Fei, T., & Shuan, L. (2018). News reliability evaluation using latent semantic analysis. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *16*(4), 1704–1711. https://doi.org/10.12928/telkomnika.v16i4.9062

Yin, R. K. (2001). *Estudo de caso: planejamento e métodos*. 2a edição. SAGE.

Yu, S., Vorobeychik, Y., Alfeld, S. (2018). *Adversarial classification on social networks*. Proceedings of the 17th international conference on autonomous agents and multiagent systems (pp. 211–219). http://dl.acm.org/citation.cfm?id=3237383.3237420

Zhang, Q., Yilmaz, E., & Liang, S. (2018). *Ranking-based method for news stance detection*. Companion Proceedings of the Web Conference 2018 (pp. 41–42). https://doi.org/10.1145/3184558.3186919

3
Avaliação Experimental

Este capítulo traz um artigo da avaliação experimental, submetido à Revista Interamericana de Comunicação Midiática - *Animus*, Qualis A3.

AVALIAÇÃO DE MÉTODOS DE MINERAÇÃO DE TEXTOS APLICADOS À DETECÇÃO DE FAKE NEWS ELEITORAIS BRASILEIRAS

Resumo

Contexto: A evolução dos meios de comunicação tem contribuído com a disseminação de notícias falsas, 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*. **Objetivo:** O objetivo deste trabalho é avaliar a eficácia dos métodos mais utilizados para verificar correspondência de textos, no contexto da detecção de notícias falsas, 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 *fakes* por seguidores de cada candidato é apresentada. **Método:** Foi planejado e executado um experimento controlado, para comparar a eficácia dos métodos selecionados. **Resultados:** 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%). **Conclusão:** 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 *fakes* por ambos os lados e que seguidores do candidato Jair Bolsonaro (PSL) foram responsáveis por 62,25% dos *tweets* relacionados a notícias falsas, 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. Notícias Falsas

Abstract

Context: 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. **Objective:** 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 fakes by followers of each candidate is presented. **Method:** A controlled experiment was planned and executed to compare the effectiveness of the selected methods. **Results:** 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%). **Conclusion:** 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 fakes 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

Contexto: 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*. **Objetivo:** 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 *fakes* por seguidores de cada candidato. **Método:** Se planificó y ejecutó un experimento controlado para comparar la efectividad de los métodos seleccionados. **Resultados:** 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%). **Conclusión:** 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 *fakes* 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. Noticias falsas

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 L e 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 notícias falsas a fatias altamente específicas da população, muitas vezes, visando gerar influência nos mais diversos segmentos da sociedade, a exemplo da política (ALLCOTT e 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 MJ, MATTHEW A, *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 (MARCO AURÉLIO, 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 e DE 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 notícias falsas, 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 *fakes* 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 *fakes* 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 notícias falsas, 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 notícias falsas, 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 e 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} ... 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 e 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 e ABUZEINA, 2017).

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

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).

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

Entretanto, como observado no trabalho de 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 e 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:

-	*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

Tabela 1 - Exemplo de cálculo do TF-IDF

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. BM25 (*Best Match 25*)

Baseado na teoria da probabilidade, o BM25 (ROBERTSON e 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.

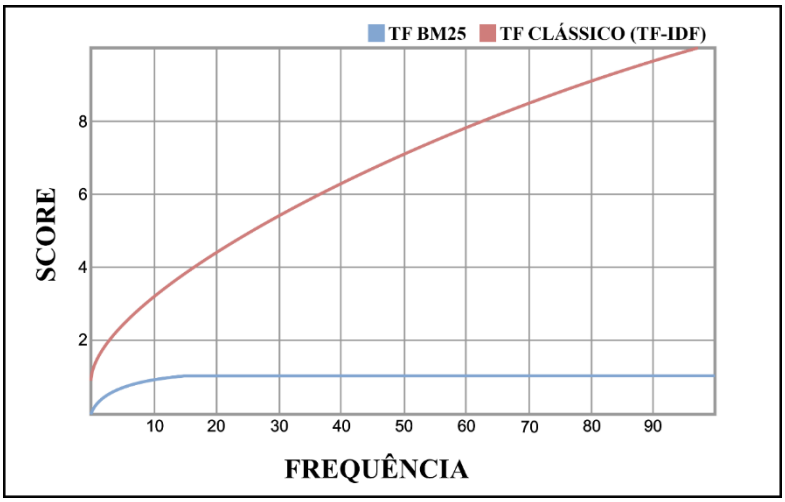

Figura 1 - Curva de saturação do componente TF (ELASTIC, 2020).

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

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

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:

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

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:

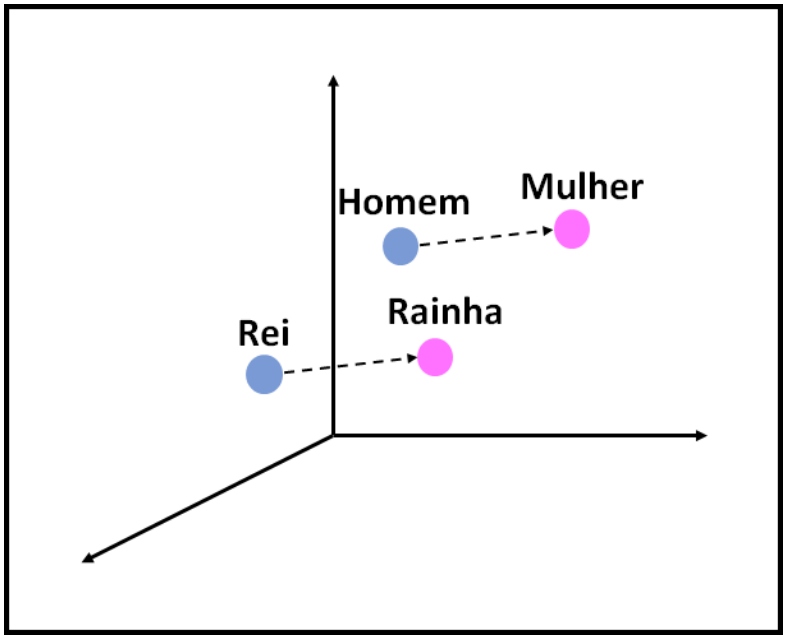

Figura 2 - Representação bidimensional simplificada para exemplificar a relação capturada entre palavras (CONTRATRES, 2020).

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 por Le e 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 e KAMBER, 2011) .

Para *n* classes, a matriz de confusão é uma tabela de dimensão *n x 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 e PATRA, 2014). As linhas correspondem às classificações corretas e as colunas representam as classificações realizadas pelo classificador (HAND, MANNILA e 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 e 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 notícias falsas (FRIGGERI, ADAMIC, *et al.*, 2014). Comparado a trabalhos existentes sobre detecção de notícias falsas com foco em eventos sociais gerais ou eventos de emergência (JIN, CAO, *et al.*, 2014), este artigo apresenta uma análise de notícias falsas, 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 e POBLETE, 2011), clientes (MORRIS, COUNTS, *et al.*, 2012), padrões de propagação (WU, YANG e ZHU, 2015), e conteúdo multimídia (JIN, CAO, *et al.*, 2015) são extraídos para que um classificador de *fakes* 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 e 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 e MEI, 2015), foi proposto um método baseado em léxico para detectar notícias falsas 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 notícias falsas. 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 *fakes*. 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.

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,

[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.

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.

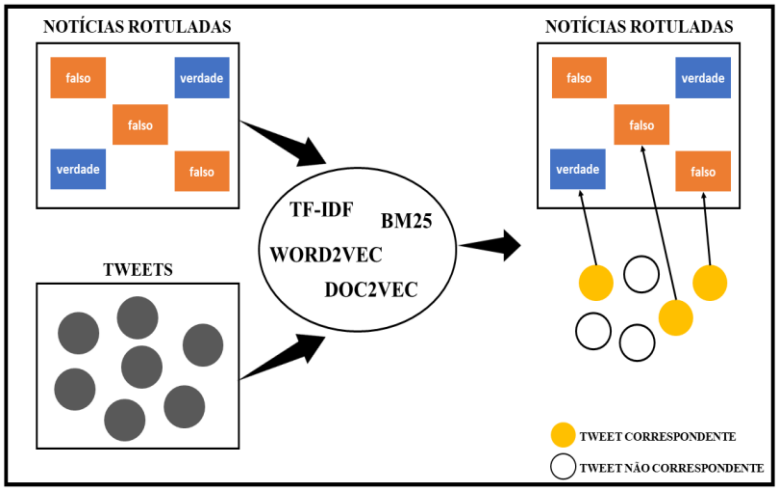

Figura 3 - Modelo de correspondência entre documentos utilizado.

Para a obtenção das métricas a serem avaliadas, utilizou-se uma adaptação da abordagem *10-Fold Cross-validation* (HASTIE, TIBSHIRANI e 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 *fakes* 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 corresponência 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 e COLAÇO JÚNIOR, 2018). A Figura 4 ilustra as etapas do trabalho, esta Seção irá focar na etapa 5.2, o planejamento do experimento.

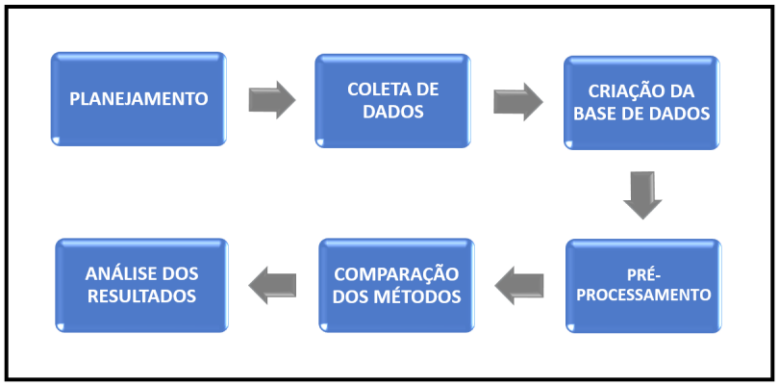

Figura 4 - Etapas do trabalho.

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 notícias falsas, avaliando e validando o que melhor se adéqua ao contexto de notícias falsas no processo eleitoral brasileiro.

Utilizando o modelo GQM (*Goal Question Metric*) (BASILI e 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 notícias falsas, **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 notícias falsas 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 notícias falsas 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 notícias falsas e 883 eram sobre fatos verídicos, como mostra a Figura 5.

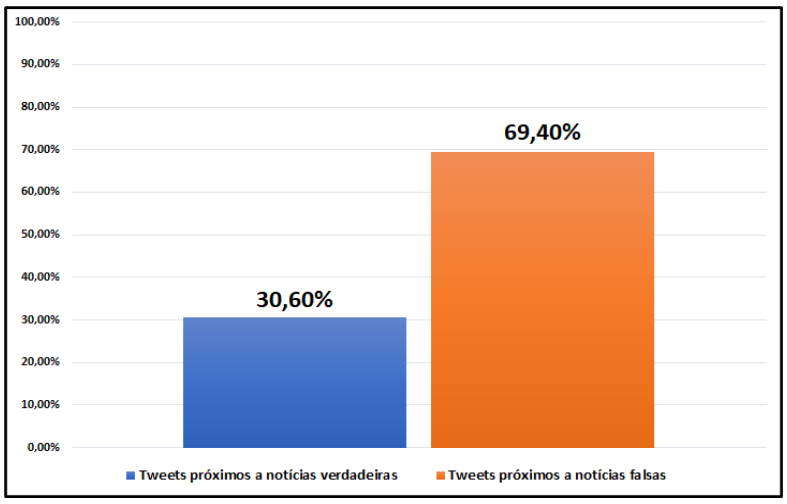

Figura 5 - Distribuição de *tweets* por rótulo.

Estes resultados colaboram com os resultados apresentados em (VOSOUGHI, ROY e 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 notícias falsas e 353 eram mais próximos de notícias verdadeiras. Entre os seguidores de Fernando Haddad, dos 1.168 *tweets*, 794 se aproximavam de notícias falsas e 374 se aproximam de notícias verdadeiras.

Como não era possível afirmar, imediatamente, que todos o tweets próximos a notícias falsas ou verdadeiras eram, respectivamente, verdadeiros positivos ou verdadeiros negativos, para que fosse gerada uma visão geral de *fakes* 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 e 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 consistiu de 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 e 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 e 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 notícias falsas checadas e de dois *tweets* antes e após essa etapa:

Notícias Falsas 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 e 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.

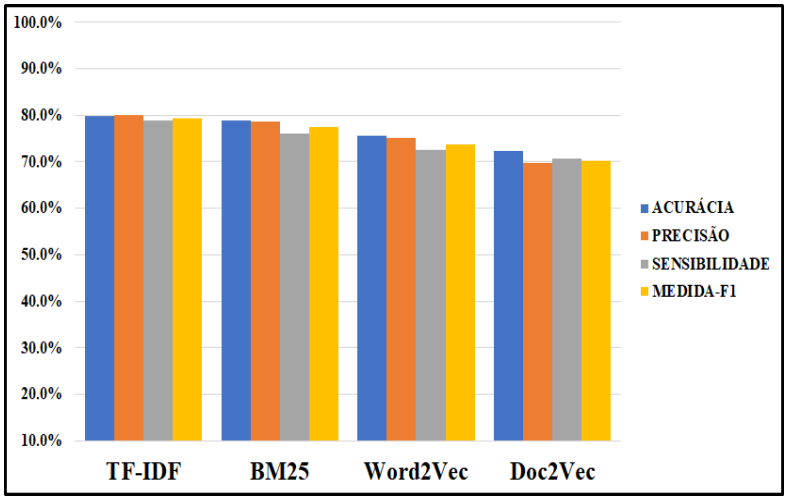

Figura 6 - Comparativo das métricas entre os métodos.

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%

Tabela 2 - Comparativo das métricas para os métodos.

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.

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

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

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.

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

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

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 e SOTO, 2013) e (DEKKER, PECHENIZKIY e VLEESHOUWERS, 2009).

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

Tabela 5 - Valores obtidos pelo Teste de Tukey para Acurácia.

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.

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
BM25-TF-IDF	0,083	0,500

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

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.

-	**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%

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

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

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 um pouco abaixo dos resultados do TF-IDF e do BM25, considerados métodos tradicionais. Desconsiderando questões de desempenho, novas pesquisas precisam analisar se os resultados deste 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, necessidade de adaptações para este novo contexto.

Partindo para uma análise mais próxima ao "negócio" eleições, os resultados demonstrados na Figura 7 fazem uma comparação da quantidade de *tweets* relacionados a notícias falsas 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 notícias falsas, contra 37,75% publicados por seguidores de Fernando Haddad. Neste sentido, mesmo com uma maior quantidade de *fakes* publicadas por parte dos seguidores do candidato do PSL, existiu disseminação de *fake news* por seguidores de ambos os candidatos.

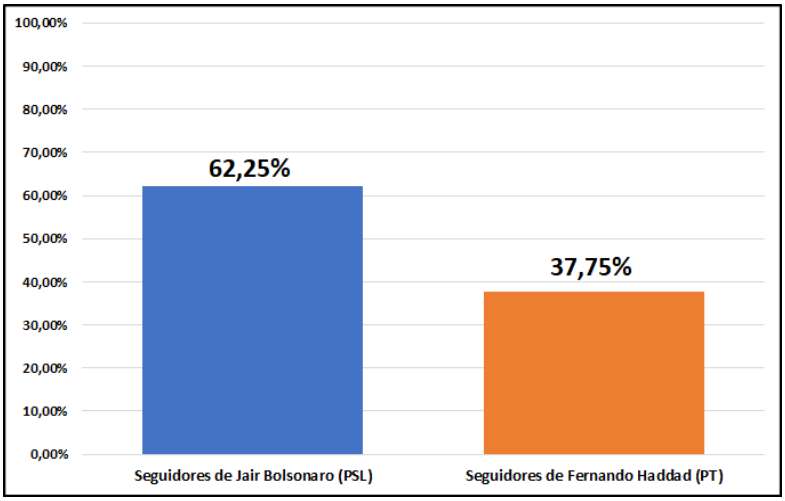

Figura 7 - Porcentagem de *tweets* sobre notícias falsas por seguidores de cada candidato.

No pleito norte-americano, também houve disseminação de *fakes* por ambos os lados, com seguidores do candidato eleito, Donald Trump, surpreendentemente, tendo publicado 18%

[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).

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 notícias falsas, como é possível observar na Figura 8.

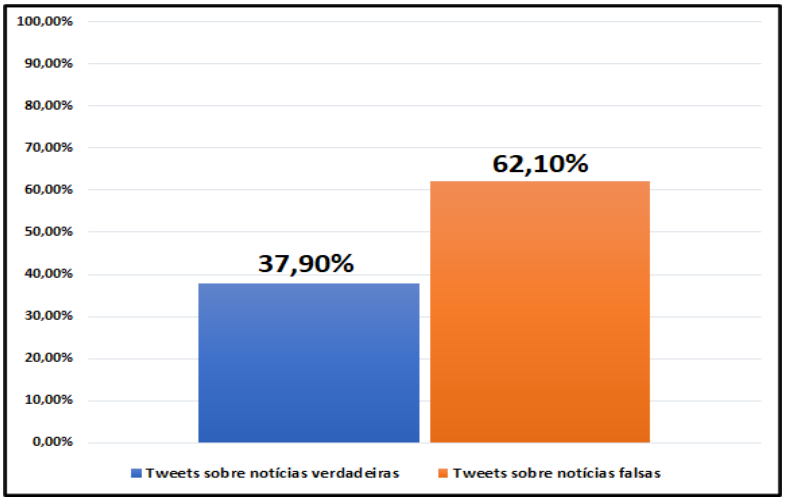

Figura 8 - Distribuição de *tweets* por rótulo dos seguidores de Fernando Haddad (PT).

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

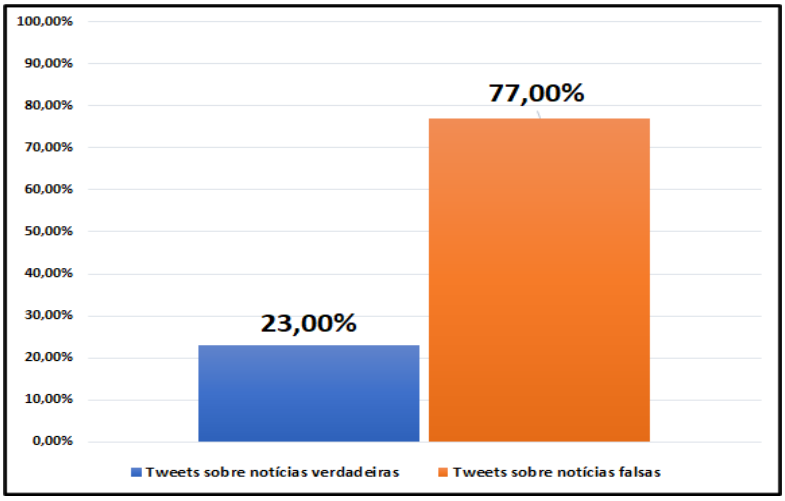

Figura 9 - Distribuição de *tweets* por rótulo dos seguidores de Jair Bolsonaro (PSL).

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.

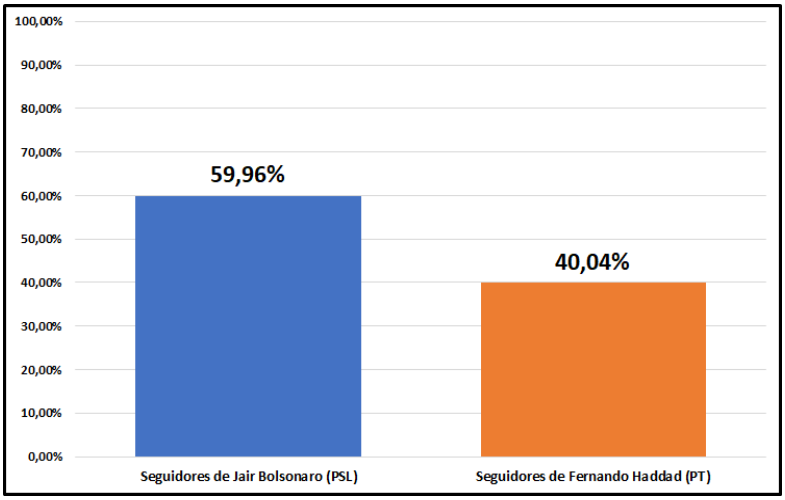

Figura 10 - Porcentagem de contas excluídas rapidamente por seguidores de cada candidato.

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 notícias falsas 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 notícias

falsas 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 notícias falsas 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*[4][5][6], três conjuntos de dados: o primeiro, (1) 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 *fakes* 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

[4] https://www.kaggle.com/caiovms/ brazilian-election-fake-news-2018
[5] https://www.kaggle.com/caiovms/brazilian-tweets-2018
[6] https://www.kaggle.com/caiovms/tweets-about-brazilian-election-fake-news-2018

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 *fakes* e que seguidores do candidato Jair Bolsonaro (PSL) foram responsáveis por 62,25% dos *tweets* relacionados a notícias falsas, 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.

9. Referências

AL-ANZI, F. S.; ABUZEINA, D. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. **Journal of King Saud University-Computer and Information Sciences**, 2017. 189-195.

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. **Journal of economic perspectives**, p. 211-36, 2017.

AMAZON. Amazon Web Services, 15 out. 2019. Disponivel em: <https://aws.amazon.com/pt/>.

ANJOS, A. Análise de Variância, 2009. Disponivel em: <http://www.est.ufpr.br/ce003/material/apostilace003.pdf>.

BASILI, V. R.; WEISS, D. M. **A methodology for collecting valid software engineering data**. [S.l.]. 1983.

BIRD, S. NLTK, 01 set. 2020. Disponivel em: <https://www.nltk.org/>.

BUCKLEY, C. **The importance of proper weighting methods**. Human Language Technology: Proceedings of a Workshop Held at Plainsboro. [S.l.]: [s.n.]. 1993. p. 21-24.

CAELEN, O. A Bayesian interpretation of the confusion matrix. **Annals of Mathematics and Artificial Intelligence**, p. 429-450, 2017.

CASTILLO, C.; MENDOZA, M.; POBLETE, B. **Information credibility on twitter**. Proceedings of the 20th international conference on world wide web. [S.l.]: ACM. 2011. p. 675-684.

CIAMPAGLIA, G. L. et al. Computational fact checking from knowledge networks. **PloS one**, p. e0128193, 2015.

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

CONROY, N. J.; VICTORIA L, R.; YIMIN, C. **Automatic deception detection:** Methods for finding fake news. Proceedings of the 78th ASIS\&T Annual Meeting: Information Science with Impact: Research in and for the Community. [S.l.]: [s.n.]. 2015. p. 82.

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

DAVID MJ, L. et al. The science of fake news. **Science**, p. 1094-1096, 2018.

DEKKER, G.; PECHENIZKIY, M.; VLEESHOUWERS, J. **Predicting Students Drop Out:** A Case Study. Proceedings of the International Conference on Educational Data Mining. [S.l.]: [s.n.]. 2009. p. 41-50.

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

FATOS, A. **Aos Fatos**, 14 ago. 2018. Disponivel em: <https://www.aosfatos.org/>.

FERNANDES, H. As novas guerras: O desafio da guerra híbrida. **Revista de Ciências Militares**, p. 13-40, 2016.

FIELD, A. **Descobrindo a estatística usando o SPSS**. [S.l.]: Porto Alegre: Artned, 2009.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the american statistical association**, 1937. 675-701.

FRIGGERI, A. et al. **Rumor Cascades**. Eighth International AAAI Conference on Weblogs and Social Media. [S.l.]: [s.n.]. 2014.

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

HAN, J.; PEI, J.; KAMBER, M. **Data Mining:** concepts and techniques. [S.l.]: Elsevier, 2011.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. [S.l.]: MIT Press, 2001.

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

JIN, Z. et al. **News Credibility Evaluation on Microblog with a Hierarchical Propagation Model**. Proceedings of the 2014 IEEE International Conference on Data Mining. [S.l.]: IEEE Computer Society. 2014. p. 230-239.

JIN, Z. et al. **News Credibility Evaluation on Microblog with a Hierarchical Propagation Model**. Proceedings - IEEE International Conference on Data Mining, ICDM. [S.l.]: Scopus. 2015. p. 230-239.

JIN, Z. 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.

LE, Q.; MIKOLOV, T. **Distributed representations of sentences and documents**. International conference on machine learning. [S.l.]: [s.n.]. 2014. p. 1188-1196.

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

LI, B.; HAN, L. **Distance weighted cosine similarity measure for text classification**. International Conference on Intelligent Data Engineering and Automated Learning. [S.l.]: Springer. 2013. p. 611-618.

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

LUPA, A. Agência Lupa, 20 out. 2019. Disponivel em: <https://piaui.folha.uol.com.br/lupa/>.

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

MAKICE, K. **Twitter API:** Up and running: Learn how to build applications with the Twitter API. [S.l.]: O'Reilly Media, Inc., 2009.

MARCO AURÉLIO, R. 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.

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

MIKOLOV, T. et al. **Distributed representations of words and phrases and their compositionality**. Advances in neural information processing systems. [S.l.]: [s.n.]. 2013. p. 3111-3119.

MONGODB. MongoDB, 15 jan. 2020. Disponivel em: <https://www.mongodb.com/>.

MORRIS, M. R. et al. **Tweeting is believing?:** understanding microblog credibility perceptions. Proceedings of the ACM 2012 conference on computer supported cooperative work. [S.l.]: ACM. 2012. p. 411-450.

MUNDSTOCK, E. et al. Introdução à Análise Estatística utilizando o SPSS 13.0. **Cadernos de Matemática e Estatística Série B**, 2006.

MYSLINSKI, L. J. US Patent 8,185,448, 2012.

OGAWA, T. Num2Words, 09 jan. 2020. Disponivel em: <https://pypi.org/project/num2words/>.

OLIVEIRA, R. A. N. D.; COLAÇO JÚNIOR, M. Experimental analysis of stemming on jurisprudential documents retrieval. **Information**, 2018. 28.

ÖZCAN, S. Tweet Pre-Processor, 01 set. 2020. Disponivel em: <https://pypi.org/project/tweet-preprocessor/>.

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

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

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

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

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

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

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

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

SANTOS, R. M. 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, L. E.; DOANE, D. P. **Estatística Aplicada à Administração e Economia**. [S.l.]: 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, E. M.; DE ALMEIDA SANTOS, J. 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. Disponivel em: <https://www.ibm.com/analytics/spss-statistics-software>.

STATISTA. Statista. **statista**, 20 mar. 2019. Disponivel em: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Acesso em: 12 Outubro 2019.

TIAN, Y.; LO, D.; SUN, C. **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, G. H.; GUROV, D.; AMARAL, E. Introdução à engenharia de software experimental, 2002.

TUMASJAN, A. et al. **Predicting elections with twitter:** What 140 characters reveal about political sentiment. Fourth international AAAI conference on weblogs and social media. [S.l.]: [s.n.]. 2010.

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

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

WANG, H. et al. **A system for real-time twitter sentiment analysis of 2012 us presidential election cycle**. Proceedings of the ACL 2012 System Demonstrations. [S.l.]: [s.n.]. 2012. p. 115-120.

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

WILCOXON, W. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, 1945. 80-83.

WOHLIN, C. et al. **Experimentation in software engineering**. [S.l.]: Springer Science & Business Media, 2012.

WU, K.; YANG, S.; ZHU, K. Q. **False rumors detection on sina weibo by propagation structures**. 2015 IEEE 31st international conference on data engineering. [S.l.]: IEEE. 2015. p. 651-662.

ZHAO, Z.; RESNICK, P.; MEI, Q. **Enquiring minds:** Early detection of rumors in social media from enquiry posts. Proceedings of the 24th International Conference on World Wide Web. [S.l.]: International World Wide Web Conferences Steering Committee. 2015. p. 1395-1405.

4
Conclusão

O objetivo principal deste estudo foi avaliar, por meio de um processo experimental, a eficácia dos métodos de mapeamento mais utilizados para a correspondência de texto, na tarefa de detecção automática de *fake news* sobre as eleições presidenciais brasileiras de 2018, comparando as evidências encontradas com os resultados obtidos de um mapeamento do estado da arte publicado nesta pesquisa.

Ao analisar os resultados para o estado da arte, objetivo secundário e basilar desta pesquisa, foi identificado que os principais algoritmos utilizados na tarefa de detecção de notícias falsas são LSTM (17,14%), Naive-Bayes e Algoritmo de Similaridade (11,43% cada um). Além disso, foi possível observar lacunas relacionadas a trabalhos no contexto *Big Data*, bem como a necessidade de replicações dos estudos existentes, na forma de experimentos mais controlados.

Após execução do experimento controlado, verificou-se que os métodos utilizados para correspondência de textos TF-IDF e BM25 obtiveram médias estatisticamente similares de acurácia, respectivamente, 79,86% e 79,00%. Também foi possível observar que os métodos *Word2Vec* e *Doc2Vec* obtiveram resultados um pouco abaixo dos demais, também respectivamente, 75,69% e 72,39%. Além disso, considerando o universo de notícias checadas disponível, o período analisado e uma margem de erro de aproximadamente 3,5%, evidenciou-se a divulgação de *fake news* da parte de seguidores de ambos os candidatos avaliados no experimento. Os seguidores do candidato Jair Bolsonaro (PSL) foram responsáveis por 62,25% dos *tweets* relacionados a notícias falsas, 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. Neste contexto, a divulgação de *fake news* nem sempre implica intenção, podendo implicar apenas um engajamento maior por parte de alguns seguidores.

Vale ressaltar que os resultados para Acurácia e Medida-F1 obtidos durante a execução do experimento foram similares aos alcançados pelo trabalho selecionado no mapeamento

da literatura. Contudo, não foi possível realizar análises comparativas em relação às métricas Sensibilidade e Precisão, pois os dados não foram disponibilizados pelos autores.

Do ponto de vista das contribuições desta pesquisa, destacam-se:

- A realização de um Mapeamento Sistemático utilizado para identificar e caracterizar as principais abordagens, técnicas e algoritmos usados, na computação, para a detecção de notícias falsas. Ressalta-se que os resultados foram publicados no *Journal of Applied Security Research*;
- A avaliação experimental de quatro métodos utilizados para verificar correspondência de textos, na tarefa de detecção automática de *fake news* sobre as eleições presidenciais brasileiras de 2018. Os resultados foram validados e submetidos à Revista Interamericana de Comunicação Midiática – *Animus*;
- A disponibilização de três *datasets* para replicações e novas análises, na plataforma *Kaggle*, contendo os seguintes conteúdos:
    1. Notícias previamente checadas sobre as eleições brasileiras de 2018;
    2. *Tweets* publicados por seguidores dos dois principais candidatos à presidência, durante o período eleitoral de 2018;
    3. *Tweets* manualmente e arduamente classificados, utilizados para averiguar a visão geral de *fake news* por seguidores.

Como consequência destas contribuições, foi respondida a questão principal de pesquisa elaborada no início do trabalho:

No contexto da detecção de notícias falsas eleitorais no *Twitter*, entre os métodos de mapeamento utilizados para correspondência de texto, selecionados no Mapeamento Sistemático da Literatura, qual o melhor em termos das métricas de qualidade a serem avaliadas (Acurácia, Precisão, Sensibilidade e Medida-F1)?

Resposta: O TF-IDF obteve os melhores resultados em todas as métricas, em comparação com os demais métodos avaliados. No entanto, após uma validação estatística, verificou-se que o TF-IDF e o BM25 obtiveram médias estatisticamente similares de acurácia, respectivamente, 79,86% e 79,00%.

Como trabalhos futuros, pretende-se avaliar os métodos de correspondência em outros contextos, uma vez que a problemática das *fake news* não é uma particularidade da política. Na saúde, por exemplo, como no caso da Covid-19, essa mesma "praga" tem se espalhado. Com um modelo mais experimentado, o investimento em uma ferramenta para checagem automática de informações torna-se mais viável.

Referências

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. *Journal of economic perspectives*, v. 31, n. 2, p. 211–36, 2017. Citado na página 8.

BASILI, V. R.; SHULL, F.; LANUBILE, F. Building knowledge through families of experiments. *IEEE Transactions on Software Engineering*, IEEE, v. 25, n. 4, p. 456–473, 1999. Citado na página 11.

CIAMPAGLIA, G. L. et al. Computational fact checking from knowledge networks. *PloS one*, Public Library of Science, v. 10, n. 6, p. e0128193, 2015. Citado na página 8.

COLLINS. *"Collins"*. 2017. <https://www.collinsdictionary.com/word-lovers-blog/new/collins-2017-word-of-the-year-shortlist,396,HCB.html>. Acessado em 25/03/2019. Citado na página 9.

CONROY, N.; RUBIN, V.; CHEN, Y. 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. Cited By 41. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84987722728&doi=10.1002%2fpra2.2015.145052010082&partnerID=40&md5=8c48a5390a854e63f2e380d1c5793223>. Citado na página 8.

DAVIS, R.; PROCTOR, C. *Fake news, real consequences: Recruiting neural networks for the fight against fake news*. [S.l.]: Stanford CS224d Deep Learning for NLP final project, 2017. Citado na página 10.

GILCHRIST, A. Post-truth: an outline review of the issues and what is being done to combat it. *Ibersid*, v. 12, n. 2, 2018. Citado na página 12.

JIN, Z. et al. Rumor detection on twitter pertaining to the 2016 us presidential election. *arXiv preprint arXiv:1701.06250*, 2017. Citado 2 vezes nas páginas 11 e 14.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007. Citado na página 10.

LAZER, D. M. et al. The science of fake news. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018. Citado na página 8.

RECUERO, R.; GRUZD, A. Cascatas de fake news políticas: um estudo de caso no twitter. *Galáxia (São Paulo)*, SciELO Brasil, n. 41, p. 31–47, 2019. Citado na página 10.

ROCHLIN, N. Fake news: belief in post-truth. *Library hi tech*, Emerald Publishing Limited, v. 35, n. 3, p. 386–392, 2017. Citado na página 9.

RUEDIGER, M. A. Robôs, redes sociais e política no brasil: estudo sobre interferências ilegítimas no debate público na web, riscos à democracia e processo eleitoral de 2018. Fundação Getúlio Vargas, 2017. Citado na página 9.

SILVA, C. V. M.; FONTES, R. S.; JÚNIOR, M. C. Intelligent fake news detection: A systematic mapping. *Journal of Applied Security Research*, Taylor & Francis, p. 1–22, 2020. Citado na página 13.

SPINELLI, E. M.; SANTOS, J. de A. Jornalismo na era da pós-verdade: fact-checking como ferramenta de combate às fake news. *Revista Observatório*, v. 4, n. 3, p. 759–782, 2018. Citado na página 9.

STATISTA. *"Statista"*. 2019. <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Acessado em 20/03/2019. Citado na página 8.

TEIXEIRA, V. M. et al. As fake news e suas consequências nocivas à sociedade. In: *Anais do Encontro Virtual de Documentação em Software Livre e Congresso Internacional de Linguagem e Tecnologia Online*. [S.l.: s.n.], 2018. v. 7, n. 1. Citado na página 12.

TRAVASSOS, G. H.; GUROV, D.; AMARAL, E. Introdução à engenharia de software experimental. UFRJ, 2002. Citado na página 11.

TWITTER. *"Twitter muda regras para combater fake news e manipulação política"*. 2018. <https://help.twitter.com/pt/rules-and-policies/twitter-report-violation>. Acessado em 20/03/2019. Citado na página 9.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1146–1151, 2018. Citado na página 12.

WHATSAPP. *"O WhatsApp continua pessoal e privado"*. 2020. <https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private>. Acessado em 09/07/2020. Citado na página 9.