Detecção de desinformação sobre Covid-19 no Twitter

Ana Alice Ximenes Mota[1], **Wellington Franco**[1] **e**
César Lincoln Cavalcante Mattos[1]

[1]Universidade Federal do Ceará – Fortaleza – CE – Brasil,

aliceximenes@alu.ufc.br, wellington@crateus.ufc.br,
cesarlincoln@dc.ufc.br

Abstract. *The damage caused by false or misleading news has increased due to the ease with which information is disseminated on social networks. During the Covid-19 pandemic, which began in 2020, such news could generate panic in the population and erroneously instruct people about the prevention of the disease. The present work introduces a new corpus from Twitter posts in the Portuguese language about misinformation from Covid-19*[1]*. In addition to the new corpus, the work evaluates different approaches to textual representations and learning algorithms models in the task of detecting misinformative messages. The best result obtained achieved an F1-score of* 89% *in the SVM classification model with the TF-IDF textual representation.*
Keywords: *Natural Language Processing, misinformation, Covid-19.*

Resumo. *Os danos causados por notícias falsas ou enganosas têm se potencializado graças à facilidade com que as informações são disseminadas em redes sociais. Durante a pandemia do Covid-19, iniciada em 2020, tais notícias foram capazes de gerar pânico na população, além de instruir erroneamente as pessoas sobre a prevenção da doença. O presente trabalho introduz um novo corpus a partir de postagens no Twitter na língua portuguesa com desinformações sobre a Covid-19*[1]*. Além do novo corpus, o trabalho avalia diferentes abordagens de representações textuais e algoritmos de aprendizagem na tarefa de detecção de mensagens contendo desinformação. O melhor resultado obtido alcançou F1-score de* 89% *no modelo de classificação SVM com a representação textual TF-IDF.*
Palavras Chaves: *Processamento de Linguagem Natural, Desinformação, Covid-19.*

1. Introdução

Um estudo publicado em parceria entre *We are social*[2] e *Hootsuite*[3], relatou que em janeiro de 2021 existiam 4,2 bilhões de usuários de redes sociais pelo mundo. Esse número equivale a um crescimento de mais de 13% em relação ao ano de 2020 e a expectativa é que esse aumento continue nos próximos anos [Kemp 2021]. Por consequência, mais pessoas e empresas consomem e expõem informações nas redes sociais. Contudo, essa liberdade acaba permitindo que ocorra uma grande disseminação de informações falsas, comumente chamadas de desinformação ou *fake news*.

[1]Disponível em https://github.com/aliceximenes/fake-news-covid-19.
[2]https://wearesocial.com
[3]https://www.hootsuite.com

De acordo com [Lazer et al. 2018], *fake news* podem ser definidas como informações fabricadas que imitam o conteúdo da mídia de notícias na forma, mas não no processo organizacional ou na intenção. A intenção, por vezes, é manipular a população em diversos contextos. Podemos citar como exemplo de manipulação no contexto político a interferência causada pela *Cambridge Analytica/Facebook* nas eleições presidenciais dos Estados Unidos em 2016 [Confessore 2018].

Com o advento da pandemia da Covid-19, ocorreu o desencadeamento de uma série de informações falsas, atingindo a população como um todo. Por exemplo, algumas desinformações buscam enganar a população indicando meios para prevenir e/ou curar a doença, colocando em risco a sociedade por se tratarem de métodos sem nenhuma comprovação científica.

Visando auxiliar a identificação de desinformações dessa natureza que possam causar malefícios à sociedade, este trabalho propõe a criação de um *corpus* em língua portuguesa sobre a Covid-19. Diversas abordagens de classificação serão avaliadas, incluindo diferentes técnicas de representação textual e algoritmos de aprendizagem de máquina, proporcionando modelos de referência para a tarefa de detecção de desinformação. O novo *corpus*, coletado da rede social Twitter[4], contém postagens com desinformação e não desinformação de cinco tópicos amplamente noticiados sobre o novo Coronavírus, causador da Covid-19.

O presente artigo é organizado da seguinte forma: na próxima seção, é feito um breve resumo dos trabalhos relacionados; na Seção 3 relata-se detalhadamente a metodologia usada para a construção do corpus, pré-processamento dos dados e modelagem da solução; na Seção 4 é feita a explicação de como os experimentos foram realizados e os resultados obtidos são discutidos; por fim na Seção 5, os resultados encontrados são discutidos e direções para investigações futuras são apontadas.

2. Trabalhos relacionados

Em [Monteiro et al. 2018], foi proposto o primeiro *corpus* na língua portuguesa com a finalidade de detectar notícias falsas. O objetivo do trabalho foi criar o *corpus*, Fake.Br, com 3100 notícias verdadeiras e 3100 notícias falsas. As notícias foram coletadas em sites e blogs, sendo as classificadas como verdadeiras extraídas de grandes veículos de notícias, como G1, Folha de São Paulo e Estadão. Em um trabalho posterior, diversas representações textuais e modelos de aprendizagem de máquina foram explorados e avaliados no mesmo *corpus* [Silva et al. 2020].

A criação do Fake.Br abriu caminho para múltiplas pesquisas científicas e para a construção de novos *corpora* sobre desinformação em língua portuguesa. Muitas dessas focam na coleta e experimentação de dados extraídos de aplicativos de mensagens e de redes sociais. Em [Cabral et al. 2021] é feita a criação do *corpus* FakeWhastApp.BR a partir de dados coletados em grupos públicos de conversas no aplicativo WhatsApp[5]. Em relação a redes sociais, uma amplamente usada e foco de pesquisas atuais é o Twitter.

Com uma projeção para o ano de 2021 de 322.4 milhões de usuários [Newberry 2021], o Twitter vem sendo alvo de diversas pesquisas acadêmicas. Em

[4]https://twitter.com

[5]https://www.whatsapp.com/

[Cordeiro and Pinheiro 2019] é feita a construção do *corpus* intitulado FakeTweet.Br com dados coletados do Twitter em português. Por se tratar de postagens que possuem limite de tamanho (280 caracteres), o *corpus* FakeTweet.Br proporciona experimentos em textos com formatos diferentes dos encontrados em blogs, sites e aplicativos de mensagens.

Na língua inglesa, em [Buntain and Golbeck 2017], foi construído um sistema automatizado para detectar desinformações em tópicos populares do Twitter. Foram utilizados dois conjuntos de dados para o aprendizado: CHEDBANK, introduzido por [Mitra and Gilbert 2015], constituindo-se de uma base de dados *crowdsourced* em que os tópicos contidos nos tweets estão identificados; e PHENE, um conjunto de dados de rumores em potencial do Twitter [Zubiaga et al. 2016]. Por fim, o método desenvolvido foi aplicado a uma base de dados de notícias falsas.

O uso de mecanismos de identificação de desinformações nos dados do Twitter também foi utilizado por [Zervopoulos et al. 2020] nos protestos políticos ocorridos em Hong Kong. Foi utilizado um conjunto inicial de postagens falsas em inglês e chinês, sendo este último traduzido para o inglês na sequência.

Uma visão geral dos trabalhos citados está apresentada na Tabela 1. Nela, são indicados o título e ano dos artigos, o idioma do *corpus*, a fonte de dados usada para sua construção e a informação se o rótulo do *corpus* foi construído de forma manual, semi-automático ou automático.

Até o conhecimento dos autores, não existia um corpus público em português obtido através de dados do Twitter com desinformações sobre a Covid-19. Diante disso, o presente trabalho, tem o intuito de construir um *corpus* com essas especificações, além de realizar experimentos e comparar os seus resultados.

Tabela 1. Resumo dos trabalhos relacionados.

Título	Idioma	Fonte	Rótulo	Ano
Automatically identifying fake news in popular twitter threads	Inglês	Sites e Twitter	Semi-automático	2017
Contributions to the study of fake news in portuguese: New corpus and automatic detection results	Português	Blogs e sites	Semi-automático	2018
Um corpus de notícias falsas do twitter e verificação automática de rumores em língua portuguesa.	Português	Twitter	Manual	2019
Towards automatically filtering fake news in portuguese	Português	Blogs e sites	Semi-automático	2020
Hong Kong Protests: Using Natural Language Processing for Fake News Detection on Twitter	Inglês	Twitter	Automático	2020
FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages	Português	WhatsApp	Manual	2021

3. Metodologia

Este trabalho propõe a criação de um novo *corpus* contendo postagens do Twitter sobre determinados tópicos falsos da Covid-19. A Figura 1 retrata as etapas realizados no fluxo de construção do novo *corpus*. Essas etapas serão detalhadas na sub-seção seguinte.

Em adição ao novo *corpus*, foi realizado um extenso trabalho de experimentação de técnicas textuais e modelos de classificação para encontrar a abordagem mais assertiva na classificação de tweets entre desinformação e não desinformação.

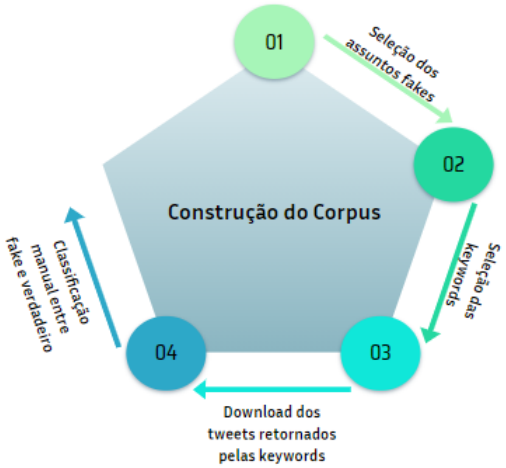

Figura 1. Fluxo de criação do corpus.

3.1. Construção do Corpus

Em meados de 2008 foi criado o primeiro *corpus* sobre desinformações na língua portuguesa [Monteiro et al. 2018] e como, até o conhecimento dos autores, não existia um *corpus* público em português, rotulado e proveniente do Twitter com dados de desinformações sobre a Covid-19. Foi necessário construir um novo *corpus* para a realização de um estudo sobre o tema.

Inicialmente, foram escolhidos tópicos falsos amplamente divulgados como métodos de tratamento e/ou prevenção eficazes contra a doença. Os tópicos estão relacionados na Tabela 2 e foram retirados do site do Governo Federal Brasileiro[6] criado para combate das desinformações sobre a Covid-19.

Tabela 2. Tópicos com desinformações usadas para coleta no Twitter.

	Tópicos
1	Chá de limão com bicarbonato quente cura coronavírus
2	Beber muita água e fazer gargarejo com água morna, sal e vinagre previne coronavírus
3	Vitamina C + zinco e o novo coronavírus
4	Ivermectina tem eficácia comprovada contra a Covid
5	Beber água quente mata o coronavírus

[6]https://antigo.saude.gov.br/fakenews/

Após a escolha dos tópicos, seguiu-se para a etapa de *web crawler* dos tweets. O processo foi realizado filtrando os tweets com as palavras chaves dos tópicos da Tabela 2 e palavras relacionadas ao novo Coronavírus. A coleta ocorreu entre fevereiro e setembro de 2020.

Posteriormente, o processo de rotulação foi iniciado. Foi realizada a leitura de cada tweet coletado e atribuído o rótulo de desinformação ou não. Para os casos que não se enquadravam nessas categorias, por exemplo por terem um viés de sarcasmo ou não serem relacionados ao assunto, o tweet não foi adicionado ao *corpus*. Também foram observados diversos tweets repetidos, os quais não foram considerados. Por fim, o *corpus* construído possui 730 tweets, sendo 456 rotulados como desinformação e 274 como não desinformação. A Tabela 3 apresenta alguns exemplos.

Tabela 3. Amostra do *corpus* criado.

Tweets	**Classificação**
O coronavírus pode ser curado se você fizer gargarejo com bicarbonato de sódio e limão.	Desinformação
Receita com limão e bicarbonato, além de não evitar mortes por coronavírus, pode ser prejudicial à saúde.	Não desinformação

3.2. Pré-processamento dos Dados

Com a construção do *corpus*, a etapa seguinte foi de tratamento e limpeza dos dados. As bibliotecas NLTK[7] e *Regex*[8] do Python foram utilizadas nessa etapa. O primeiro passo foi colocar todos os tweets em letras minúsculas. Além disso, foram retirados os acentos, pontuações, números, palavras de parada (*stopwords*) e eventuais links que direcionavam para outros sites. Em seguida, foi realizado o processo de tokenização, que consiste em dividir os textos em segmentos demarcados de caracteres, chamados *tokens*.

Por fim, experimentamos algumas técnicas de codificação textual amplamente usadas na literatura. Foram testadas as seguintes técnicas: o tradicional método de *Bag of Words* (BoW) e Frequência do Termo – Frequência Inversa dos Documentos (TFIDF); e o método de representação textual Word2Vec [Mikolov et al. 2013].

As codificações BoW e TFIDF padrão tratam cada palavra de forma individual, descartando inteiramente a ordem de aparecimento no texto. A técnica *ngram* permite preservar algumas dessas informações. Essa abordagem permite considerar uma sequência de palavras adjacentes como um único termo. Com isso, a sequência pode ser composta por uma palavra (forma padrão do BoW e TFIDF) ou *ngramas*, em que *n* é a quantidade de palavras adjacentes.

Dessa forma, para as representações BoW e TFIDF foram testados diferentes valores para o hiperparâmetro *ngram_range* (1gram, 2gram ou 3gram) e foi definido a remoção de termos com uma frequência menor ou igual a 1% nos tweets coletados. Na representação Word2Vec, usou-se os vetores de palavras pré-treinados propostos em [Hartmann et al. 2017]. Os modelos utilizados para gerar esses vetores foram treinados

[7]https://www.nltk.org/

[8]https://pypi.org/project/regex/

com documentos da língua portuguesa de 17 conjuntos de dados de diferentes domínios, totalizando 1.395.926.282 *tokens*. Foram considerados vetores com 300 dimensões treinados com a abordagem *Skip-Gram*, pois foi a abordagem com o melhor resultado experimental obtido da técnica.

3.3. Modelos de Aprendizagem de Máquina e Métricas de Avaliação

Tendo finalizado o pré-processamento dos dados, iniciou-se o preparo do *corpus* para a etapa de aprendizagem. Para tanto, dividiu-se a base de dados aleatoriamente em 70% para treino e 30% para teste. A última parcela foi usada exclusivamente para avaliar a capacidade de generalização dos algoritmos de aprendizagem.

Os modelos de classificação binária comparados foram: LightGBM (LGBM), Máquina de Vetores de Suporte (*Support Vector Machines*, SVM), *Random Forest* (RF), *Naive Bayes* (NB), AdaBoost (AB) e Regressão Logística. Para o LGBM, foi usada a implementação original[9], para os demais modelos foram usadas as implementações da biblioteca scikit-learn [Pedregosa et al. 2011][10].

Para cada modelo de aprendizagem avaliado foi realizado um processo de escolha dos melhores hiperparâmetros usando o método de pesquisa em grade (*grid search*), implementado na biblioteca scikit-learn. Para validar o resultado do modelo em cada combinação de hiperparâmetros foi realizado, na base de treino, o cálculo da métrica F1-score e o método de validação cruzada em *k-folds*, sendo $k = 5$.

As métricas consideradas para a avaliação final dos classificadores, nos dados reservados para teste, foram as seguintes: acurácia, indica a performance geral do modelo, ou seja, a fração de quantos tweets com desinformação foram preditos como desinformação; precisão, a fração dentre todas as classificações de tweets em desinformação pelo modelo, quantas estão corretas; revocação, dentre todos os casos de tweets com desinformação qual a fração de acerto; F1-score é a média harmônica entre precisão e revocação; e a área sob a curva ROC, indica a capacidade discriminativa do modelo, isto é, a capacidade de classificar corretamente tweets em desinformação e não desinformação.

4. Experimentos e Discussão de Resultados

Considerando-se a parcela de dados reservada para treinamento (70% do *corpus*), a etapa de *grid search* para a escolha da melhor combinação de hiperparâmetros dos modelos envolveu também a avaliação de diferentes *pipelines* de processamento do *corpus*. As *pipelines* consideradas testaram as codificações BoW e TFIDF com diferentes valores para o hiperparâmetro *ngram*, além de avaliar a representação textual Word2Vec. A Tabela 4, apresenta os hiperparâmetros testados em cada modelo. Ressalta-se que os códigos dos experimentos e o *corpus* construído estão disponíveis em repositório público.

Os resultados obtidos com as representações BoW e TFIDF aplicadas na base de teste, estão apresentados na Tabela 5. Destaca-se que os modelos obtidos proporcionam soluções de referência para pesquisas futuras. Nota-se que em todos os modelos as métricas foram, em sua maioria, acima de 80%, indicando bons resultados de detecção.

[9]https://lightgbm.readthedocs.io/

[10]https://scikit-learn.org/

Tabela 4. Hiperparâmetros testados em cada modelo.

Modelos	Hiperparâmetros
LGBM	Número máximo de folhas em uma árvore e regularização L1 e L2
SVM	Função de kernel e gamma dependendo da função de kernel
Random Forest	Número de estimadores e índice de pureza
Naive Bayes	Parâmetro de suavização aditivo
AdaBoost	Número de estimadores, taxa de aprendizagem e algoritmo utilizado
Reg. Logística	Tipo de penalização e o algoritmo utilizado

Tabela 5. Melhores resultados de cada modelo (BoW e TF-IDF).

Modelo	Acurácia	Precisão	Revocação	F1-Score
LGBM Classifier	0,78	0,80	0,83	0,82
SVM	**0,85**	**0,82**	**0,97**	**0,89**
Random Forest	0,84	0,83	0,91	0,87
Naive Bayes	0,79	0,87	0,77	0,81
AdaBoost	0,84	0,86	0,89	0,87
Reg. Logística	0,82	0,79	0,95	0,86

Na Figura 2, temos a curva ROC e a área sob a curva ROC (AUC, *area under the curve*) indicando bons resultados em todos os modelos, em especial o Máquina de Vetores de Suporte (SVM). A métrica usada para escolher o melhor modelo foi o F1-score. Logo, o melhor modelo obtido nesse experimento foi o Máquina de Vetores de Suporte com 89% de F1-score. Para esse modelo, a melhor *pipeline* e conjunto de hiperparâmetros encontrados foi utilizando a codificação TF-IDF, *1gram*, usando a função de kernel RBF e o *gamma scale*.

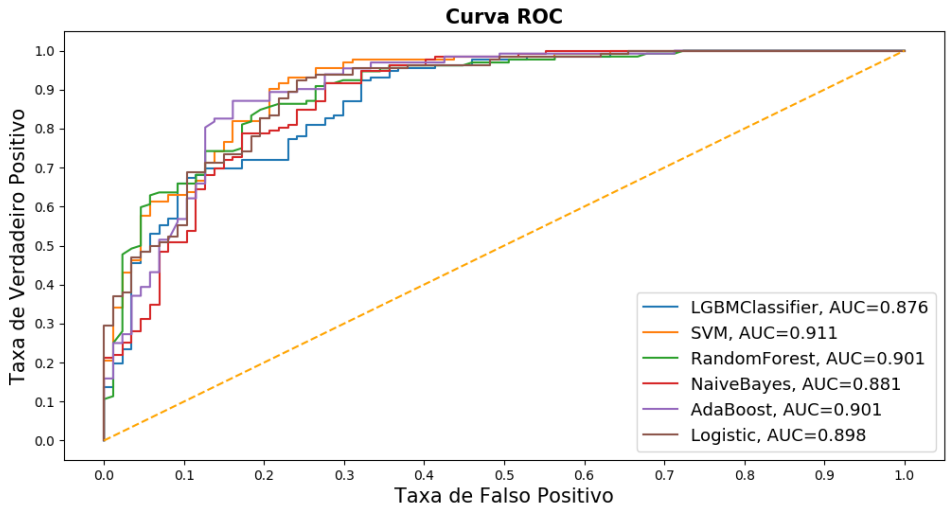

Figura 2. Curva ROC dos experimentos com BoW e TFIDF.

Na Tabela 6 e Figura 3 estão os resultados dos experimentos com a técnica Word2Vec. O desempenho foi abaixo do encontrado na abordagem anterior. O comporta-

mento médio das métricas foi em torno de 70% e não houve um modelo que se sobressaiu aos demais. Entretanto, entre todos os modelos, o LGBM e o SVM foram os que obtiveram maior F1-score, cerca de 76%. O LGBM obteve o melhor desempenho, com $76,8\%$ de F1-score. Para esse modelo, o conjunto de hiperparâmetros escolhido foi o valor 5 para o número máximo de folhas, a regularização L1 com termo $0{,}1$ e sem regularização L2 (termo igual a 0).

Tabela 6. Melhores resultados de cada modelo com Word2Vec.

Modelo	Acurácia	Precisão	Revocação	F1-Score
LGBMClassifier	**0,69**	**0,70**	**0,84**	**0,76**
SVM	0,61	0,61	1,00	0,75
RandomForest	0,66	0,66	0,88	0,76
AdaBoost	0,67	0,69	0,81	0,75
Reg. Logística	0,62	0,62	0,92	0,74

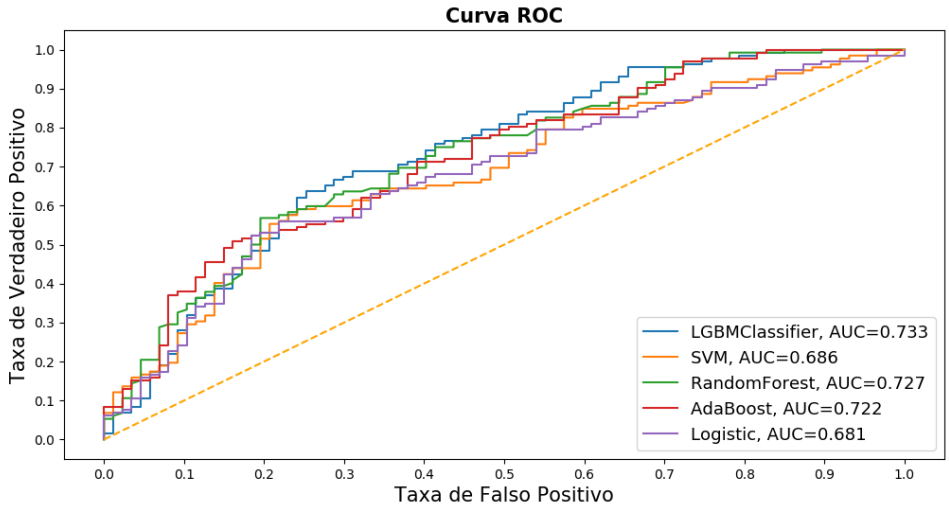

Figura 3. Curva ROC dos experimentos com Word2Vec.

5. Conclusão

O presente trabalho propôs a construção de um *corpus* coletado do Twitter com desinformações relacionadas à Covid-19. O novo *corpus* foi usado em experimentos que avaliaram diferentes representações textuais e modelos de aprendizagem de máquina para classificar tweets em desinformação ou não desinformação. Os resultados indicaram o modelo SVM com representação textual TFIDF como a melhor abordagem, com F1-score de 89% usando *1gram* e função de *kernel* RBF.

Trabalhos futuros envolvem a análise de erros dos classificadores. Identificando se os modelos erram ou não os mesmos tweets, se os tweets mais difíceis de classificar corretamente possuem pontos em comum, se um modelo de *ensemble* com os classificadores que cometem erros distintos resulta em um melhor resultado. Outra vertente consiste em investigar questões éticas relacionadas a eventuais vieses na classificação.

Referências

Buntain, C. and Golbeck, J. (2017). Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE.

Cabral, L., Monteiro, J. M., da Silva, J. W. F., Mattos, C. L. C., and Mourao, P. J. C. (2021). FakeWhastApp.BR: NLP and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 63–74. INSTICC, SciTePress.

Confessore, N. (2018). Cambridge analytica and facebook: The scandal and the fallout so far. `https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html`. Acessado em : 20/07/2021.

Cordeiro, P. R. and Pinheiro, V. (2019). Um corpus de notícias falsas do twitter e verificação automática de rumores em lingua portuguesa. In *STIL-Brazilian Symposium in Information and Human Language Technology. IEEE, Salvaldor, BA, Brazil*, pages 220–228.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Kemp, S. (2021). Digital 2021: the latest insights into the 'state of digital'. `https://wearesocial.com/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital`. Acessado em : 20/07/2021.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitra, T. and Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth international AAAI conference on web and social media*.

Monteiro, R. A., Santos, R. L., Pardo, T. A., De Almeida, T. A., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *International Conference on Computational Processing of the Portuguese Language*, pages 324–334. Springer.

Newberry, C. (2021). 36 twitter statistics all marketers should know in 2021. `https://blog.hootsuite.com/twitter-statistics/`. Acessado em : 20/07/2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Zervopoulos, A., Alvanou, A. G., Bezas, K., Papamichail, A., Maragoudakis, M., and Kermanidis, K. (2020). Hong kong protests: using natural language processing for fake news detection on twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 408–419. Springer.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., and Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.