Classifying Fake News: Analysis of Trends and Correlations in Real and Fake Articles

Authors: Dalton Gilmore, Yahia Mohamed, Jongkook Son, Caleb Van Tassel

Abstract -**The spread of the internet and news platforms has made it very easy for any individual to be able to reach a large audience. In this environment, misinformation can spread at a very rapid rate, so there is a need for a systematic way of determining if an article is portraying true or false information. With the development of machine learning techniques such as Natural Language Processing, we are now able to analyze text computationally, and derive trends that exist in both real and fake articles. In this paper, we present some of these analyses performed on a real/fake news dataset to obtain a better idea of what characteristics are unique to each type of article. We perform topic modeling, sentiment analysis, and ngram decomposition to extract further information about the articles. Also, being able to pull out these trends computationally is reason to believe that a neural network could be trained to pick up these trends via supervised learning and classify the articles automatically. To demonstrate this idea, we train multiple types of neural network models with the dataset, and obtain a classification accuracy of 98.6%. This serves as evidence that there are linguistic characteristics of falsified articles that are detectable by a neural network.**

Introduction

Since before the emergence of the internet, publishers have used falsified information to further their interests. The dissemination of misinformation and fake news has become extremely fast during this era of advancement of technology and availability of the internet, social media, and different media channels. The users have access to a variety of web and online platforms which made the reach to fake publications more convenient and faster. Additionally, clickbait has been used by publishers to disseminate misinformation and fake news [1]. During the time of elections, the amount of misinformation and fake news increases significantly as the different political parties use the available media platforms for their agenda and to attract more voters. Additionally, the high level of misinformation that spread during the current COVID-19 pandemic and the vaccination process led to a dangerous health issue crisis and cost many lives.

The degree of fake news and misinformation spread limits the ability of journalists to differentiate between news and information and this task becomes troublesome and time-consuming. In response to this issue, researchers and media experts have developed a fake news detector that adopts natural language processing (NLP) to analyze the patterns of the word and find the statistical correlations among the news articles [2]. Also, there are some organizations such as PoliticalFact.com, and FacktCheck.org trying to address the issue of political fake news during the election time using the fact-checking method and investigating comments of any organizations, public figures, and journalists [3]-[4]. The giant tech companies such as Twitter, Facebook, and Google have different attempts to address and fight the spread of fake news and misinformation. However, the easy accessibility of these platforms by millions of users and their sharing option made the dissemination of these fake news and misinformation worse[1].

Although extensive research has been done in this area, still there is a need to find a more effective method for identifying the type of news and information in a fast and efficient manner. We propose that falsified information often has characteristics that indicate its validity, such as the use of extreme (strongly negative or positive) language, the topics which are discussed, and other, less intuitive, characteristics. In this paper, we will use a machine learning approach to identify and compare trends that exist in fake news articles, and show that these can be identified computationally using a classification network.

Related Work

Allcott and Gentzkow [5] discussed the theoretical and empirical background frame for the debate whether the spread of fake news helped Donald Trump to be elected as president of the United States in the 2016 election or not. Also, they explained the effect of the economics of fake news, then presented data on the consumption of fake news by the public before the election. Additionally, they compared the importance and effect of social media platforms to other sources of political news and information. By using statistical reports, they provided more information on the rate at which voters were affected due to the exposure to fake news. According to their database, Donald Trump widely shared fake stories. They used a survey to study inference of news headlines and discussed the factors that were highly related to the consumption of fake news. They found that the consumption of fake news articles was high by average US adults during the election period which may change people's vote depending on the effectiveness of fake news. Still, this analysis has limited ability to decide whether the fake news was a strong factor that helped Donald Trump to win the 2016 election or not.

Tsfati et al.[6] conducted a literature review to find out the role of mainstream news media in spreading fake news, why they give more coverage for this fake news, and the effect of these coverages on their audiences. According to their analysis, the mainstream news media have a significant role in spreading fake news. There were two reasons that these news media cover fake news; first, because of massive news values and these media platforms trying to find the truth. Second, the partisan and ideology of the news media. Also, they found this coverage has a significant effect on their audiences. Although the research in this area is growing fast, still, there are many limitations such as the exact definition of fake news, the type of stories that receive bigger attention in the media coverage, and the styles of reporting of these fake news that are still not thoroughly investigated.

Mesquita et al.[7] discussed the spread of fake medical news and explored their effect during the COVID-19 pandemic. Fake medical news can mislead others in order to damage organizations or make profits. They brought up one of the most famous examples of misinformation in public health, in which the misconception that measles, mumps, rubella (MMR) vaccine causes autism was created by a fraudulent article published in one of the top medical journals (Lancet). Due to the fast spread of this misinformation around the world, measles outbreaks started to appear in different countries. Also, they discussed the available tools to fight this misinformation by using artificial intelligence applications such as machine learning and natural language processing methods. Additionally, the vital strategy is to provide credible, evidence-based information to the public through liable organizations (i.e. WHO, OPAS, etc.).

Treharne and Papanikitas [8] defined fake news and discussed the tools used to detect and report it in the medical and health field. Also, they showed the difference between fake news and poor reporting in the health and medical field. They defined fake news as “fabricated information that mimics news media content in form but not in organizational process or intent”. However, they note, the reader needs to differentiate between entirely fabricated news items and inaccurate or misleading news items. Additionally, they explained the areas of validation that can be used to detect entirely fabricated news items such as original sources, the scope of coverage, use of different fact-checking sites, and use of publication titles to generate more search results on the topic. They explored five areas of inaccurate and sufficiently misleading news items that commonly appear:contextualization, causality, risk, extrapolation, credibility. Therefore, these should be checked to detect inaccurate or misleading news.

Rashkin et al. [3] conducted an analytic study of the language of news media and political quotes which was written with different intents and levels of truth. Satire, hoax, and propaganda samples were compared to find any linguistic characteristics for untruthful text. Also, they presented a case study using PoliticalFact.com and their factuality judgments on a 6-point scale to perform fact-checked statements and determine their text truthfulness. They concluded that the task of fact-checking still challenging and understating the lexical features of the text help in discrimination between the reliable and unreliable sources of digital news.

Vargo et al.[4] investigated the relationship between fake news, fact-checkers, and online news media by focusing on partisan media using the computational approach for the online new media landscape from 2014 to 2016. They found that there is influence on agendas of partisan and coverage of breaking news by fake news. They also pointed out the election in 2016 shows how the partisan media was susceptible to the agenda of fake news. The fact-checkers faced different difficulties such as the determination of the agenda of the news media, which could limit its ability to stop the spread of fake news and dissemination of their corrections. One of the major challenges of this study was the measurement of the agenda-setting power of false claims that fake news spread.

Zhou et al. [2] conducted an experiment to show the vulnerability of fake news detection methods that use only NLP. They found that the methods that lack semantic knowledge were associated with a high false-positive rate which led to misclassification of real news. They evaluated a model called Fakebox on three classes of adversarial attacks that focus on various aspects of every article such as fact distortion, subject-object exchange, and cause confounding. These adversarial examples of real news were generated from a dataset for fake and real news known as McIntire’s dataset. They found this model was performed poorly and its accuracy was significantly lowered with these adversarial attacks. They proposed the use of fact-based knowledge should be adopted with NLP-based models such as crowdsourced knowledge graph to overcome this drawback of these detectors.

Oshikawa et al. [9] systematically reviewed and compared the task formulations, datasets and then summarized the NLP approaches and results to detect the fake news. Still, binary classification or regression are the widely used methods for fake news detection. Although the classification is used more frequently, news that is partially real and fake requires a better approach to solve this problem. Adding additional classes is a commonly used practice to address this issue. With the regression method, the challenge is to find the right method to convert discrete labels to numerical scores. They categorize the public fake-news datasets into three main categories: entire articles (e.g Bs Detector), claims (e.g PolitiFact), and Social Networking Services (e.g BuzzFeedNews). After preprocessing steps for input text, they used the hand-crafted features extraction (rhetorical approach) for a dataset that includes the whole article length. Collecting evidence was used for all dataset that has evidence and compared the results of classifications of these data using different machine learning models. They focused on three datasets (Liar, Fever, and FakeNews Net). The results showed on the Liar dataset, as the tendency LSTM models achieved higher accuracy compared with CNN models. Also, attention-LSTM has a better score in both verification and evidence-collection tasks. There are some critiques to these methods such as the hand-crafted features are essential to non-neural network approaches but can be replaced by neural networks.

Silva et al. [10] searched libraries to find the most recent papers on fake news detection in social media and mapped the state of the art of fake news detection, defining fake news, and finding the best machine learning method of detecting it. They found that the most used method to automatically detect fake news is an amalgamation of techniques coordinated by a neural network rather than a classical machine learning approach. Also, they identified the significant need for using the ontology that will help in the unification of different definitions and terminologies used in the fake news domain.

Aldwairi and Alwahedi [1] proposed an approach that allows users to detect and filter clickbait by installing a tool into their personal browser. They crawled the web to collect URLs for the clickbait and focused on social media websites such as Facebook, Forex, and Reddit. The Python script is used to compute the attributes from the title and the content of these websites. Lastly, they extracted the features such as keywords, titles from the web pages. This tool shows a strong performance in identification of the fake news sources. The limitation of this tool will be the difficulty to discriminate pages that have titles and content written in a professional style and still include fake news or misinformation.

Dataset Description

The dataset that we chose for analysis in accord with our hypothesis is a collection of both fake news and real news from Kaggle [11]. When it comes to defining what is “real” and “fake,” the dataset’s real articles are from a reliable source and the fake articles are from a variety of unreliable sources. More specifically, the real articles are directly from an international news agency company by the name of Reuters, and the fake news includes Wikipedia pages and articles from unreliable websites as determined by Politifact, a fact-checking website that rates source accuracy. Each set of articles is contained in a comma-separated values file, labeled “Fake.csv” and “True.csv”. A majority of these articles focus on political and world news, sharing either real or biased information on events all around the world. One concern that we considered in particular when approaching this dataset was the oversampling of political and world news sources coupled with an undersampling of local news sources. We consider this alongside other assumptions as we work with the data.

Within the dataset are columns labeled “title”, “text”, “subject”, and “date”. We only utilize the columns of title and text, as the date and subject of the articles do not have an effect on whether they are true or not. Both sets are read in from their respective csv files, with the “truth” variable of the fake news set being set to 0 and the “truth” of the real news set being set to 1.

The articles themselves can easily be considered contemporary, even though most of them were created in 2016 or 2017. It is unlikely that the common strategies for falsifying and spreading false information has gravely changed since then, so the problem of outdated information or analyses should not be present in the implementation of our hypothesis.

As far as assumptions go, the only important one concerning our data is that the news data is indeed from the United States. With this assumption, we confirm that our findings related to this dataset are relatively local and might not necessarily be able to be applied to areas outside of the U.S.

Overall, this dataset was perfect for testing our hypothesis that certain aspects of how an article is written, such as the presence of strong positive or negative language, will often be key in ascertaining the validity of the article. It is through multiple data analysis methods that we employ this prediction on our dataset. This collection of both truthful and malicious news is a huge first step towards the ultimate, hopeful goal that one day we will be able to pass an article through an algorithm that will faithfully tell us what to believe.

Analysis

Data Pre-Processing

The preprocessing of the data can be broken down into four main steps:

> **Cleaning the text**: this process includes reading the text from the training and testing file and then removing the punctuation and any special character and making sure that all the text is in lower character. We also remove the beginning part of the

text which starts with ‘Reuter’s ‘ and then remove the stop word in english which .

Train/Validation Split: In this process, we split the training text into train and validation text using the Scikit learning model. We split the dataset into 70% train text, and 30 % validate the text.

Tokenization: Tokenization is the process of breaking complex data like paragraphs into simple units called tokens. These tokens are individual words that help make the bag of words and other steps for machine learning models.

Word Embedding: Word embedding is also called vectorization. In this process, we are going to use CountVectorizer to vectorize words in the text and convert the text data to numerical vectors. These steps are also called feature extraction from the text and map them to a real number vector.

Sentiment analysis

There are two datasets one contains not fake news and the other is fake news, but this was not sufficient to train the sentiment analysis model and the real problem was that we could not get the labels of sentiment from the dataset. So instead of creating labels for the sentiment, We decided to adopt another approach to use the vader model . VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data.

VADER sentiment analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

VADER’s SentimentIntensityAnalyzer method takes in a string and returns a dictionary of scores in each of four categories. And we can display polarity scores based on that for our combined dataset negative score is 0.222, the neutral score is 0.499, a positive score is 0.279. We can find out that based on the valder model, most of the text is categorized as neutral.

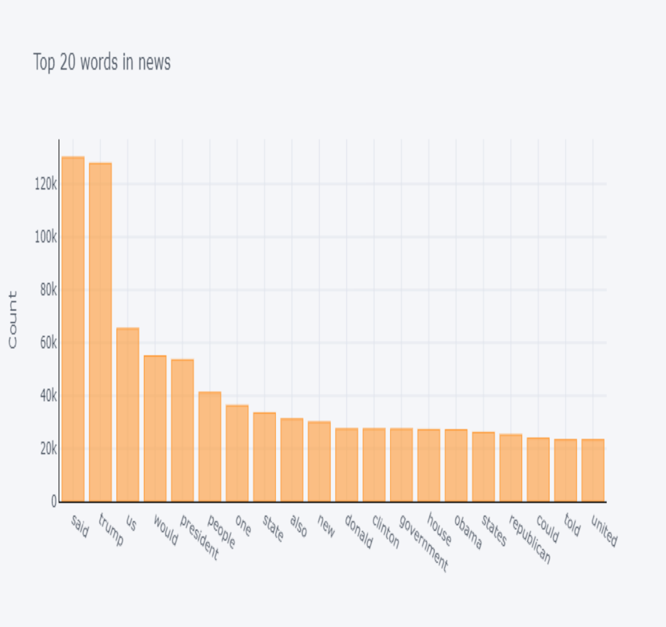
Ngram Analysis

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words, or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. Here We are going to extract words from a combined dataset that contains both True and fake news.

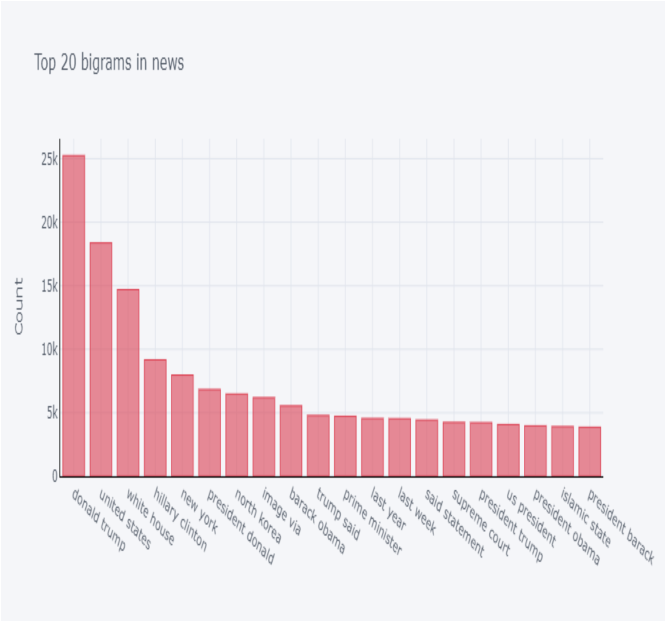
First, Let's look at the top 20 words from the news which could give us a brief idea of what news is popular in our dataset. All the top 20 news are about the US government Especially it's about Trump and the US followed by Obama. We can understand that the news is from Reuters which contains fake news.

Also, You can figure out the bigram result from fig6. As feared, the most bigram You can find out is Donald Trump. I think the model will be biased in its results considering the amount of trump news. We can see the North Korea news as well, I think it will be about the dispute between the US and North Korea. There is also few news from fox news as well.

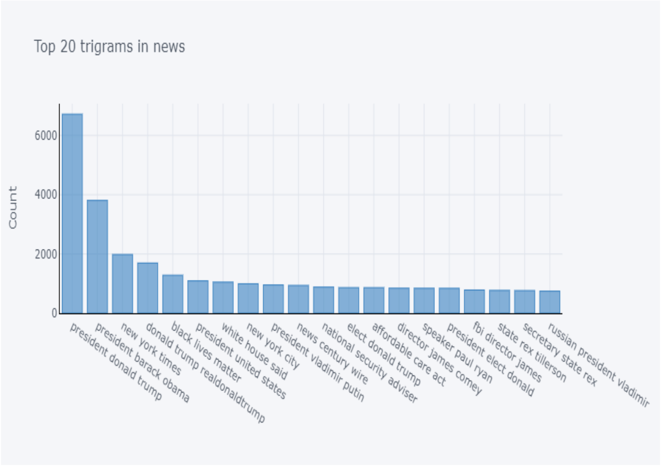
Finally, Let’s look at Trigram results in figy7. There is no big difference with bigram results. You can find out that most trigrams are related to Trump or Obama. One distinctive thing is that there is important news that ruled the US media-'Black lives matter' post the demise of Floyd. We can see that news has been covered in our data. There was a lot of fake news that revolved around death. Rest of the news is about US politics



<Fig4 Top 20 words>

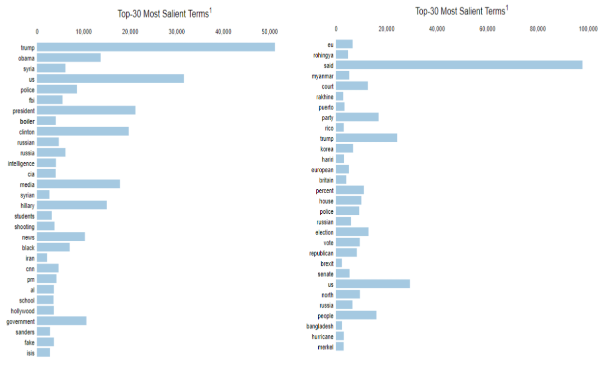


<Fig5 Top 20 bigrams>



<Fig6 Top 20 trigrams >

Topic Modeling



Fake

Real

<Fig7 Salient Terms>

Topic modeling analysis aims to identify the most prominent topics that exist within the dataset. We suspected that the topics discussed in the real and fake datasets would differ, so we ran our analysis on both datasets. As you can see from Figure 8, the topics did, in fact, differ. It appears that topics discussed in the fake dataset usually pertain to more controversial topics, such as election candidates, social issues, and government agencies. In the real dataset, there is a much greater representation of more globalized topics, such as countries that haven't been at war with the US, European issues like Brexit, and natural disasters.

From this analysis, it appears that fake news articles focus on problems that are associated with fear and slander. This makes sense, as those are likely the things that not only help to get clicks but are also the types of things that an individual party might want to skew opinions on to further their agenda. From a classification perspective, it is likely that a model could pick up on this difference computationally, and be able to correctly determine which articles are valid or not.

Implementation

Validity Classification

Given that we were able to recognize trends exclusive to each dataset in our analysis, it seemed plausible that a neural network could learn to discern whether any given article contained those trends, and thereby classify if the information was likely to be valid or not. The hope is that through unsupervised training, a network would be able to pick up on the correlations explored above, and perhaps even some less intuitive ones that we may have missed. Regardless, simply showing that these correlations exist is reasoning to believe that there are features of the data to be learned.

Implementation Details

Seeing as we desired a binary classification of valid or not, and required that the model be able to learn from both categorical and continuous data, a logistic regression model seemed like a logical choice of architecture.

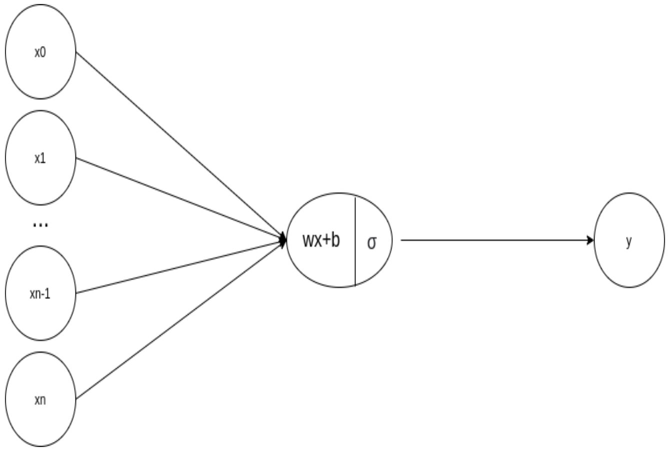
We chose to implement our model in Tensorflow, using the Scikit Learn library for preprocessing, training, and evaluation. The training process ran for 500 iterations using L2 normalization.

Preprocessing

Before we could train our model, we needed to preprocess the input such that the network could properly transform it between layers. Since our articles' subject matter was consistent throughout both datasets, we decided to employ the NLP strategy of count vectorization to convert our corpus into a vectorized representation. After removing the common stop-words, count vectorization yielded a simple representation of the articles that is computationally friendly to a neural network.

With our only categorical feature having only two classes, we simply used a label encoding method to assign 0 values to world news and 1 values to political news articles.

Logistic Regression Model



For our purposes, SKLearn's Logistic Regression model was adequate. Since our dataset was dense and not too large, we used the Limited-memory Broyden-Fletcher-Goldfarb-Shanno optimization algorithm in the training process.

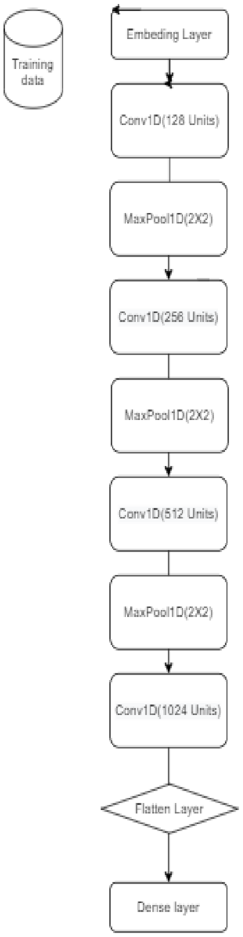
CNN Model

In deep learning, we generally use Convolutional Neural Networks and their variants to classify image data. So most people think we can use them only for image data. But a convolution operator extracts features from the data given. And if data has dimensions of more than one, we can use it with a convolution operator. And if we use word embeddings to convert words we can use a Convolutional Neural Network.

To make sure we were covering our bases, we also went ahead and implemented a CNN for our dataset.

The first input to the model is the embedding layer. Then there is a convolutional layer with filter size 128, which is to extract the text features after that batch normalization layer is added to make the artificial neural network faster and stable through normalization of the input layer by recentering and rescaling. The third layer is the max-pooling layer, which computes the region's maximum area by dividing the input into rectangular pooling regions. In our model, We set the max pooling area to be 2x2. This process is repeated three more times by increasing the unit 256, 512, 1024 with nine layers, and at the end, there is a flattened layer and dense layer which connect all the neurons.

Validation is an important part of any model, and it gives assurance that the model is working well for the dataset. By validating the model, you can evaluate the performance of the model for the given dataset.



<Fig 1 CNN model>

Evaluation

As we suspected, the models were able to pick up on the features of each dataset. With our Logistic Regression model, we were able to obtain an accuracy of 98.6% in classifying which articles were likely to come from the fake dataset. This is just slightly better than CNN, which obtained an accuracy of 91 %.

While it seems very good, this score is actually much greater than we expected. We suspect that this could be due to the discrepancy of sources in each dataset. In the fake dataset, the articles are sourced from a variety of different places. This helps to diversify the writing style and format of the data. However, in the real dataset, all of the articles are sourced from one website: Reuters. It is possible that the network simply learned the editing style of Reuter's and classified it based on that. If we had sourced our own dataset (rather than grabbing one off of Kaggle), it would've been more fair to vary the sources of the articles in the real dataset.

Conclusion

Overall this analysis and modeling shows promise for being able to accurately and automatically detect when an article contains false information. Our work could be extended to include some sort of publicly available ontology for determining common knowledge information like who is currently president. Even without this, it is clear that when falsifying information, there are frequent indicators throughout the language used.

References

[1] M. Aldwairi and A. Alwahedi, "Detecting Fake News in Social Media Networks," *Procedia Comput. Sci.*, vol. 141, pp. 215–222, Jan. 2018, doi: 10.1016/j.procs.2018.10.171.

[2] Z. Zhou, H. Guan, M. M. Bhat, and J. Hsu, "Fake News Detection via NLP is Vulnerable to Adversarial Attacks," *Proc. 11th Int. Conf. Agents Artif. Intell.*, pp. 794–800, 2019, doi: 10.5220/0007566307940800.

[3] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 2017, pp. 2931–2937, doi: 10.18653/v1/D17-1317.

[4] C. Vargo, L. Guo, and M. A. Amazeen, "The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016," *New Media Soc*, 2018, doi: 10.1177/1461444817712086.

[5] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[6] Y. Tsfati, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, and E. Lindgren, "Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis," *Ann. Int. Commun. Assoc.*, vol. 44, no. 2, pp. 157–173, Apr. 2020, doi: 10.1080/23808985.2020.1759443.

[7] C. T. Mesquita *et al.*, "Infodemia, Fake News and Medicine: Science and The Quest for Truth," *Int. J. Cardiovasc. Sci.*, vol. 33, no. 3, pp. 203–205, May 2020, doi: 10.36660/ijcs.20200073.

[8] T. Treharne and A. Papanikitas, "Defining and detecting fake news in health and medicine reporting," *J. R. Soc. Med.*, vol. 113, no. 8, pp. 302–305, Aug. 2020, doi: 10.1177/0141076820907062.

[9] R. Oshikawa, J. Qian, and W. Y. Wang, "A Survey on Natural Language Processing for Fake News Detection," *ArXiv181100770 Cs*, Mar. 2020, Accessed: Apr. 22, 2021. [Online]. Available: http://arxiv.org/abs/1811.00770.

[10] F. C. D. da Silva, R. Vieira, and A. C. Garcia, "Can Machines Learn to Detect Fake News? A Survey Focused on Social Media," 2019, doi: 10.24251/HICSS.2019.332.

[11] C. Bisaillon, "Fake and real news dataset," 26-Mar-2020. [Online]. Available: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset. [Accessed: 01-May-2021].