_____

A Review and Analysis on Fake News DetectionBased on Artificial Intelligence and Data Science

Shoaib Shaik,

Software Engineer,

Department of Computer science, Eastern Illinois university,

Charleston IL,

shoaibshaik828@gmail.com

Abstract

Along with the development of the Internet, the emergence and widespread adoption of the social mediaconcept have changed the way news is formed and published. News has become faster, less costly andeasily accessible with social media. This change has come along with some disadvantages as well. In particular, beguiling content, such as fake news made by social media users, is becoming increasingly dangerous. The fake news problem, despite being introduced for the first time very recently, has becomean important research topic due to the high content of social media. Writing fake comments and news on social media is easy for users. This study attempts to investigate advanced and state-ofthe-art fake news detection mechanisms pensively. We begin with highlighting the fake news consequences. The objective of this study was to synthesize the current possibilities of using AI to combat disinformation. Disinformation is currently a growing problem, and it is increasingly important to understand how we can utilize AI to help address this issue. Fake news and disinformation (FNaD) are increasingly being circulated through various online and social networking platforms, causing widespread disruptions andinfluencing decision-making perceptions. Despite the growing importance of detecting fake news in politics, relatively limited research efforts have been made to develop artificial intelligence (AI) and machine learning (ML) and Natural Language processing (NLP) oriented FNaD detection models suited to minimize supply chain disruptions (SCDs). Using a combination of AI and ML, and case studies this study made a review analysis by considering the authors their own perceptions in the previous years was presented in this paper.

Keywords: Fake news detection, Artificial intelligence, Machine learning, data science, Natural Language processing (NLP)., Salivary Gland Cytopathology (MRSSGC), Risk of Malignancy, FNAC, Cytology, Histopathology.

1. INTRODUCTION

The increased scholarly focus has been directed to fake news detection given their widespread impact on supply chain disruptions, as was the case with the COVID-19 vaccine. Fake news and misinformation are highly disruptive, which create uncertainty and disruptions not only in society but also in business operations. Fake news and disinformation-related problems are exacerbated due to therise of social media sites. Regarding this, using artificial intelligence (AI) to counteract the spread of false information is vital in acting against disruptive effects (Gupta et al., 2021). The era of smartphones and social media has revolutionized the way content is transmitted and received, but not all of this content is accurate or truthful. The widespread use of the Internet and social media platforms has led toa significant increase in the prevalence of misleading information, known as disinformation. It is easilyaccessible and spreads rapidly in these digital environments. And the model for Fake News Detection is shown in figure 1.
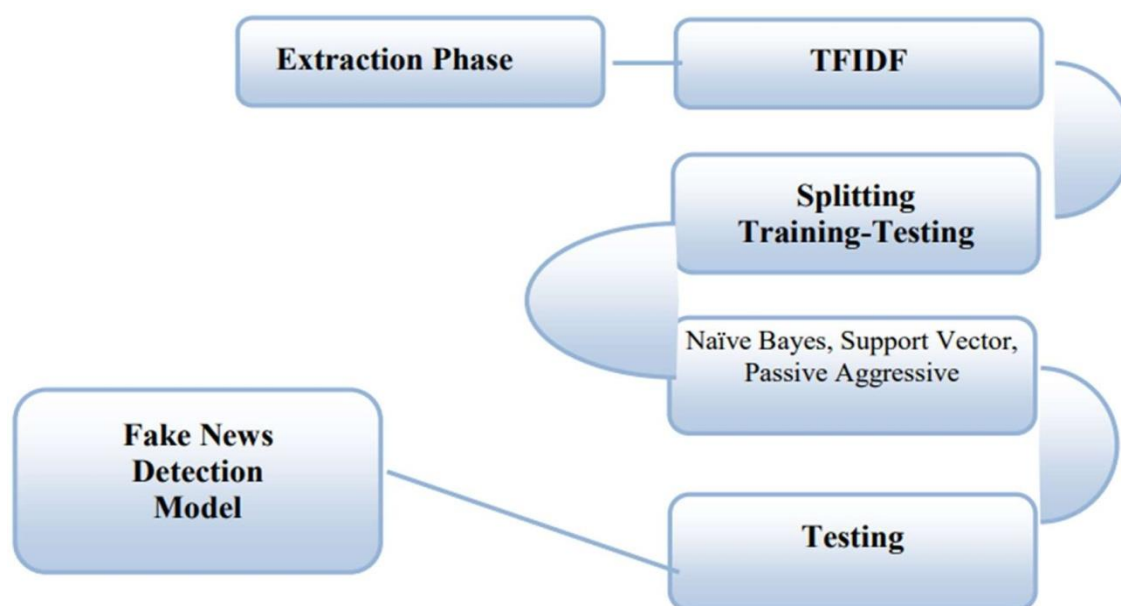
_____

Fig. 1 Fake News Detection Model

In this article, we consider the definition of disinformation from the European Commission: "false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit" (**De Cock Buning 2018, p. 10**). Disinformation, disguised as factual information, creates a distorted understanding of reality, leading to severe consequences for society by distorting people's perceptions of various issues. Political and health matters are particularly vulnerable to disinformation, as seen during the COVID-19 pandemic, during which false information spread rapidly through social media. However, disinformation can affect various areas by disseminating false knowledge about reality.

Although the phenomenon of disinformation gained significant attention during the 2016 US election campaigns, it has taken on a new dimension since 2020. The COVID-19 pandemic highlighted the critical importance of reliable and fact-based information for decision-making across all aspects of society, emphasizing its role as a fundamental pillar of democracy (**Grizzle et al. 2021**).

Disinformation not only poses challenges to discerning truthful information but also undermines journalistic credibility. It represents a significant obstacle to journalism as a means of knowledge production in society. Technological advancements, especially in artificial intelligence (AI), have not only increased the spread of disinformation but have also facilitated its automated creation and distribution. Consequently, the use of AI in disinformation creation and dissemination poses a significant challenge to the reliability of information, making it increasingly difficult to distinguish between facts and falsehoods.

AI can be defined as the process of automating tasks that typically require human intelligence. In other words, it involves replicating diverse facets of human thinking and behavior. However, **Bostrom** (**2018**) emphasized that the most significant challenge for AI lies in attaining this behavior when performing tasks that necessitate common sense and comprehension of natural language, areas in which humans instinctively excel without conscious effort.

AI systems strive to replicate human-like reasoning, learning, planning, creativity, and numerous other capabilities. Algorithms, which comprise sequences of instructions or operations, are crafted to accomplish specific objectives. Present endeavors in the development of systems to identify misleading language in disinformation have emphasized automatic classification techniques and a range of algorithms (**Bhutani et al. 2019**).

_____

One notable subfield of AI is machine learning, which enables computers to autonomously identify patterns in extensive datasets through algorithms without the need for explicit programming. Machine learning plays a vital role in tasks that require machines to learn from experience (**Oliveira 2018**). Consequently, it focuses on creating algorithms and procedures to enhance computers' performance in executing activities.

Furthermore, while AI can be used to amplify disinformation, it also plays a crucial role in identifying disinformation. According to **Singh et al.** (**2021**), advancements in this domain involve multimodal automated detection, which incorporates both textual and visual cues to identify disinformation. Various machine learning and classification algorithms have been employed to assign categories to specific datasets based on their characteristics.

2. REVIEW OF LITERATURE

Fake news, disinformation, and supply chain disruptions

Fake news (Oxford English Dictionary, 2021a) is defined as: "*false reports of events, written and read on websites.*" Furthermore, disinformation is construed as: "*false information that is given deliberately*". The impact of FNaD is substantial, disrupting economic operations and societal activities.FNaD also threaten brand names and potentially affect the consumption of products and services, ultimately impacting supply chain operations and demand. These are affected by panic-driven or bad decisions based on disinformation.

SCD is defined as a disturbance in the flows of material, financial, and information resources betweenfirms and their major stakeholders—e.g., suppliers, manufacturers, distributors, retailers, and customers. Disruption may affect supply chain operations for random periods (Mehrotra & Schmidt, 2021). Supply chains encompass the activities needed for firms to deliver products and services to their final consumers, and accurate information is an integral part of such chains, as it enablesdecision-makers to make decisions on future demand, supply, cash flows, returns, among other supplychain operations. There are historical examples of how FNaD can affect the supply chain and business operations. From the 1950s to 1990, the tobacco industry constantly shared disinformation on the adverse effects of active and secondhand smoke exposure by manipulating research, data, and the media.

In September 2006, the Royal Society, Britain's premier scientific academy, wrote to ExxonMobil urging it to stop funding the dozens of groups spreading disinformation on global warming and claiming that the global temperature rise was not related to increases in carbon dioxide levels in the atmosphere.In 2013, the Associated Press official Twitter account was hacked and a tweet was made about two explosions injuring President Barack Obama; within hours, this wiped US$130billion from the stock market (Parsons, 2020), which affected stock supply chain operations. In 2017, six UK Indian restaurants fell victim to fake news stories claiming that they were serving human flesh. One restauranthad to cut staff hours and saw its revenue fall by half. Such events could also have indirect effects on supply chain operations, ultimately being conducive to SCDs.

Artificial Intelligence in Automated Detection of Disinformation:

One challenge associated with utilizing AI and machine learning for disinformation detection constitutes ethical concerns. Biases and prejudices can infiltrate these systems, leading to erroneous outcomes. Moreover, ethical considerations pose obstacles to the use of AI in journalism, including thelack of monitoring and transparency and the potential suppression of creativity. Nevertheless, as statedby Jamil (2021), AI has been widely employed in journalism to automate repetitive tasks, particularly in data collection and identifying relevant patterns and trends for news reporting.

World is changing rapidly. No doubt we have a number of advantages of this digital world but it also has its disadvantages as well. There are different issues in this digital world. One of them is fake news.

_____

Someone can easily spread a fake news. Fake news is spread to harm the reputation of a person or an organization. It can be a propaganda against someone that can be a political party or an organization. There are different online platforms where the person can spread the fake news. This includes the Facebook, Twitter etc. Machine learning is the part of artificial intelligence that helps in making the systems that can learn and perform different actions (Donepudi, 2019). A variety of machine learning algorithms are available that include the supervised, unsupervised, reinforcement machine learning algorithms. The algorithms first have to be trained with a data set called train data set. After the training, these algorithms can be used to perform different tasks. Machine learning is using in different sectors to perform different tasks. Most of the time machine learning algorithms are used for prediction purpose or to detect something that is hidden.

Online platforms are helpful for the users because they can easily access a news. But the problem is this gives the opportunity to the cyber criminals to spread a fake news through these platforms. This news can be proved harmful to a person or society. Readers read the news and start believing it without its verification. Detecting the fake news is a big challenge because it is not an easy task (Shu et al., 2017). If the fake news is not detected early then the people can spread it to others and all the people will start believing it. Individuals, organizations or political parties can be effected through the fake news. People opinions and their decisions are affected by the fake news in the US election of 2016.

Different researchers are working for the detection of fake news. The use of Machine learning is proving helpful in this regard. Researchers are using different algorithms to detect the false news. Researchers in (Zhou et al., 2019) said that fake news detection is big challenge. They have used the machine learning for detecting fake news. They found that the fake news are increasing with the passage of time. That is why there is a need to detect fake news. The algorithms of machine learning are trained to fulfill this purpose. Machine learning algorithms will detect the fake news automatically once they have trained.

3. FAKE NEWS DETECTION USING NLP

Machine learning algorithms have been found to be effective in identifying fake news by analyzing patterns and features in text, images, and videos. One of the most widely used techniques is natural language processing (NLP), which involves analyzing text data to identify patterns and features that can indicate whether a piece of news is fake or not. NLP techniques, such as sentiment analysis, topic modeling, and text classification, have been used to detect fake news by analyzing the language, sentiment, and topic of news articles. One study used a supervised machine learning approach to classify news articles as real or fake, using a dataset of news articles from various sources. The study found that the algorithm achieved an accuracy of 92.5% in identifying fake news articles, and that the most important features for identifying fake news were the source, the language, and the sentiment of the article. Another study used a combination of NLP and deep learning techniques to detect fake news on social media platforms, achieving an accuracy of 96.2% (Mr. Alok Mishra 2023).

Fake news detection research is still in the early stage as this is a relatively new phenomenon in the interest raised by society. Machine learning helps to solve complex problems and to build AI systems nowadays and especially in those cases where we have tacit knowledge or the knowledge that is not known. We used machine learning algorithms and for identification of fake news; we applied three classifiers; Passive Aggressive, Naïve Bayes, and Support Vector Machine. Simple classification is not completely correct in fake news detection because classification methods are not specialized for fake news. With the integration of machine learning and text-based processing, we can detect fake news and build classifiers that can classify the news data. Text classification mainly focuses on extracting various features of text and after that incorporating those features into classification. The big challenge in this area is the lack of an efficient way to differentiate between fake and non-fake due to the unavailability of corpora. We applied three different machine learning classifiers on two publicly available datasets. Experimental analysis based on the existing dataset indicates a very encouraging and improved performance (Sajjad Ahmed, 2020).

_____

TABLE I SEVEN TYPES OF FAKE NEWS

| Sr. No	Type	Details
1.	False Connection	When headlines, visuals or captions don't support the content.
2.	False Context	When genuine content is sharedwith false contextual information
3.	Manipulated Content	When genuine information or imagery is manipulated to deceive.
4.	Satire	No intention to cause harms but has potential to fool.
5.	* Misleading Content	Use of information to frame an issue.
6.	Imposter Content	When genuine sources are impersonated
7.	Fabricated Content	New content that is 100% false, designed to deceive and doharm.
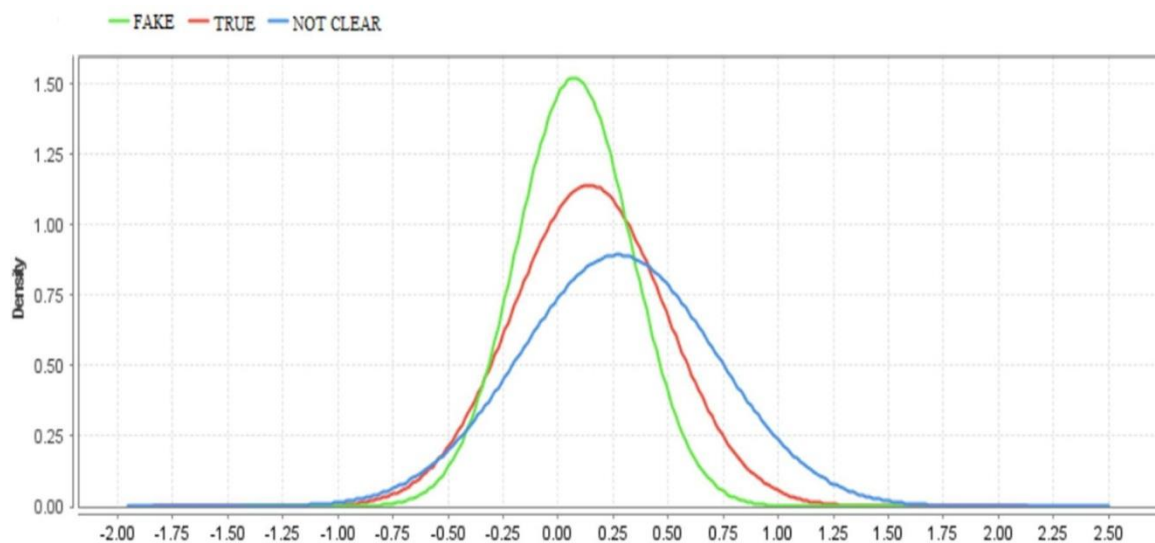
Fig. 2 Dataset class labeling chart

We used RapidMiner, a powerful machine learning tool for data exploration, preparation, information extraction, result visualization and result optimization. We analyzed the fake and true sentences through RapidMiner and initial results can be seen in Fig. 2.

_____

5. DATA SCIENCE APPLICATION ON DETECTION OF FAKE NEWS

Initially, the platform must be constructed in accordance with the data format associated with false and authentic news. The implemented programmes must be synchronised with the data structure during the design phase. The bogus database displays no news channel names, but the genuine dataset displays individual headquarters for each station. Manipulating the concept of dataset fraudulent channels are exploiting an unregistered news portal. As a result, using the original dataset, one may compare and explicitly identify them. In this venture, we are using LS-TM Recurrent Neural Network using (Long Short Term Memory) to forecast fake news because there is a large amount of fake news in all types of media such as social media or news media, and the author is training LS-TM' Genuine' and 'Fake' news data were used to train a neural network. We found FAKE NEWS messages on Twitter on the internet(Sapna B Kulkarni 2022).

Two key factors need to be considered for the receiver to accurately detect fake news: effective news representative features and appropriate model design. Despite this, many of the current solutions rely solely on content-based features, which are inadequate and can overlap. Additionally, most of the models used for classification are based on the concept of dense feature vectors, which are unsuitable for short news sentences. To overcome this issue, this study introduces a Web-Informed-Augmented Fake News Detection Model using Stacked Layers of Convolutional Neural Networks and Deep Autoencoder (ICNN-AEN-DM). The contribution of this study lies in the development of a novel Web-Informed Augmented Fake News Detection Model, which addresses the limitations of existing models by enriching content-based features with additional features gathered from trusted sources through a web search. The augmented information thereby gathered will either support or reject the claims in the news content. Then, a stacked layer of CNN with a deep autoencoder is trained using probabilistic deep learning. The probabilistic outputs of the stacked layer are used to train decision-making stacked layers based on multilayer perceptron (MLP). After conducting numerous experiments on difficult datasets, it was demonstrated that the model proposed in this study outperformed the state-of-the-art models (**Abdullah Marish Ali** 2023).

Fake news has emerged as a major challenge for journalism, democracy, and the freedom of expression. Its impact on public trust in government is a cause for concern, and its potential influence on critical events like the 2016 U.S. presidential election and the disputed "Brexit" vote remains unclear. The issue has been exacerbated by social media platforms like Facebook and Twitter, which provide users with fast, inexpensive, and less regulated information than traditional news channels like newspapers and television. According to the study, fake news on Twitter often receives much more user retweets than true news and spreads much more quickly, especially when it comes to political news. During the 2016

U.S. presidential election, the top 20 fabricated stories shared on Facebook generated over eight million shares, reactions, and comments. False information not only exacerbates division and polarization but also deepens the divide as individuals and groups become more entrenched in their own beliefs, unwill- ing to engage in constructive discourse with those who have differing perspectives (Nistor, A.; 2022). Furthermore, fake news can have serious consequences for public health and safety, such as spreading misinformation about COVID-19 or promoting unproven treatments. Therefore, it is crucial to effectively detect and prevent the spreading of fake news.

To gain some insights about the features that were extracted from the original news dataset and the features extracted using the augmented dataset gleaned from a web search, the term cloud-based analysis has been considered. The difference between density and size variance can give insights into the performance of the feature extraction and learning. **Figure 11** shows how the distribution of the features changed, based on the source of the information. **Figure 11**a presents the word cloud of the original dataset while **Figure 11**b presents the word cloud of the augmented dataset (web search features have been included). The word cloud of the original dataset is sparse compared with the word cloud of the augmented dataset, which has more features. By combining these two text features, more informative features were extracted, resulting in improved classification performance of the proposed model.
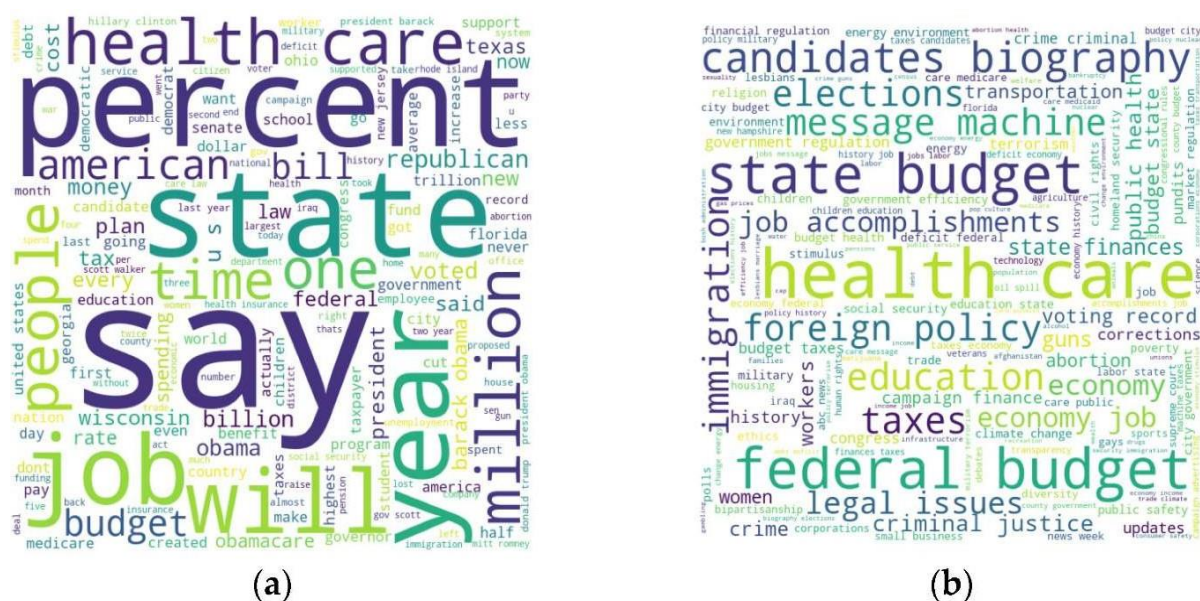
_____

Figure 11. Words important within (**a**) the original dataset (**b**) the augmented dataset.

To sum up, it is a difficult undertaking to accurately detect classified fake news when relying only on the text features extracted from news articles due to their overlapping nature with true news and lack of sufficient distinguishing features. Extracting news features is also costly and subject to various types of inaccuracies and falsehoods, such as those stemming from political or financial motivations. Researchers have tried to improve text classification accuracy by developing more advanced features based on syntax, semantics, and contextual information. However, to our knowledge, extracting similar information from reliable sources has not yet been investigated in the literature. This study shows that searching for similar information (e.g., facts) in reliable news sources boosts the performance of the classification. Features representation and model design also play important roles in the performance of the detection model. For example, TF/IDF with the Ensemble classifier achieved higher accuracy than the combined GloVe with CNN as reported by previous studies (Ali, A.M.; 2022).

Fake news refers to the manufactured information created to distort people's understanding of events. The spread of fake information has made specialists from different domains determined to implement fake news detection methods to identify and mark information that aims to affect people's understanding of reality.

In this context, Natural Language Processing (NLP) has become an important technique for the fake news detection process. NLP is a subfield of Artificial Intelligence that allows computers to understand and process human languages. Over recent years, NLP has proved its efficiency for multiple tasks such as text summarization, question answering, machine translation, text classification, etc. (Zhou, M.; 2021). Unfortunately, there is one big challenge that frequently occurs in the context of certain NLP tasks involving not-so-common languages— the lack of large and reliable data sets. To overcome this problem, augmentation techniques for text data have gained interest. Data augmentation is a well-known method for improving Machine Learning (ML) models by creating data out of data; the scope is to increase the quantity and diversity of training data sets [Vu, D.T.; 2022]. Text data augmentation is the process of creating new text data that are similar to the original text information. Increasing the size of data sets is important for NLP tasks because it allows ML models to increase their robustness.

One of the common problems in the field of NLP tasks when considering languages that are less commonly used is the lack of reliable data sets. This is a well-known challenge that has a negative impact on training the supervised ML algorithms and implicitly affects the process of making accurate predictions when it comes to NLP tasks. Another challenge in this area is related to the size, origin, and reliability of the data. Another point

_____

involves the difficulties that researchers encounter when trying to identify sources of annotated data that align with their specific problems and can be used further for data set creation. Annotated data are challenging to identify and represent the key to obtaining good results from supervised learning techniques. The most common approach to data set creation involves careful analysis of online data sources to address certain problems, rigorous data collection processes, and complex preprocessing steps to ensure that the data are suitable for applying computational algorithms. Unfortunately, these steps do not fully overcome this common challenge. Usually, more techniques are required to obtain larger data sets that can be used to train ML algorithms. The size of the data sets can be increased by using text data augmentation methods, such as translation, back translation, synonym word replacement, generative models, random insertion, swapping, or deletion. These methods have been successfully proven to enrich training data sets, and, furthermore, they can solve other common problems, such as data imbalance, by generating additional samples for the minority output class.

CONCLUSION

In conclusion, this study discusses a long-term strategy to develop models for predicting fake news before it spreads and to research the effectiveness of personalized intervention mechanisms. The combination of AI and machine learning technologies, such as in the practice of fact-checking, encompasses the three stages of fact-checking: identification, verification, and distribution the identification stage gathers the highest concentration of technologies. This study also focuses on short news sentences where the news content lacks sufficient features for learning. For longer news content, more investigations are required to collect the augmented features through a web search. The long news content may be split into smaller sentences by applying the proposed web information-collecting algorithm on each sentence to obtain probabilistic pre-decision. The final decision about the news class can be inferred by aggregating the results of the pre-decision.

REFERENCES

[1] Gupta, A., Li, H., Farnoush, A., & Jiang, K. (2021). W. Understanding Patterns of COVID Infodemic: A Systematic and Pragmatic Approach to Curb Fake News. Journal of Business Research.

[2] De Cock Buning, Madeleine. 2018. *A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation*. Luxembourg: Publications Office of the European Union.

[3] Grizzle, Alton, Carolyb Wilson, Ramon Tuazon, Chi Kim Cheung, Jesus Lau, Rachel Fischer, Dorothy Gordon, Kwame Akyempong, Jagtar Singh, Paul R. Carr, and et al. 2021. *Media and Information Literate Citizens: Think Critically, Click Wisely!* Paris: UNESCO Publishing.

[4] Bostrom, Nick. 2018. *Superinteligência*. Rio de Janeiro: DarkSide Books.

[5] Bhutani, Bhavika, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake News Detection Using Sentiment Analysis. Paper presented at the 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, August 8–10; pp. 1–5

[6] Oliveira, Arlindo. 2018. *Mentes Digitais*, 2nd ed. Lisboa: IST Press.

[7] Singh, Vivek K., Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology* 72: 3–17.

[8] Oxford English Dictionary (2020a). Oxford, UK: Oxford University Press. https://www.oxfordlearnersdictionaries.com/definition/english/fake-news

[9] Mehrotra, M., & Schmidt, W. (2021). The value of supply chain disruption duration information. *Production and Operations Management*, *30*(9), 3015–3035

[10] Parsons, D. D. (2020). The impact of fake news on company value: evidence from tesla and galena biopharma. *Chancellor's Honors Program Projects.* https://trace.tennessee.edu/utk_chanhonoproj/2328

[11] Jamil, Sadia. 2021. Artificial intelligence and journalistic practice: The crossroads of obstacles and opportunities for the Pakistani journalists. *Journalism Practice* 15: 1400–22.

[12] Donepudi, P. K. (2019). Automation and Machine Learning in Transforming the Financial Industry. Asian Business Review, 9(3), 129-138. https://doi.org/10.18034/abr.v9i3.494

[13] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

[14] Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake News: Fundamental theories, detection

_____

strategies and challenges. In WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining (pp. 836-837). (WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining). Association for Computing Machinery, Inc. https://doi.org/10.1145/3289600.3291382

[15] Mr. Alok Mishra, "FAKE NEWS DETECTION USING MACHINE LEARNING: A REVIEW", 2023 IJCRT | Volume 11, Issue 1 January 2023 | ISSN: 2320-2882

[16] Sajjad Ahmed, "Development of Fake News Model Using Machine Learning through Natural Language Processing", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:14, No:12, 2020

[17] Sapna B Kulkarni, "DETECTION OF FAKE NEWS THROUGH IMPLEMENTATION OF DATA SCIENCE APPLICATION", International Journal of Research Publication and Reviews, Vol 3, Issue 7, pp 3862-3868, July 2022

[18] Abdullah Marish Ali, "Web-Informed-Augmented Fake News Detection Model Using Stacked Layers of Convolutional Neural Network and Deep Autoencoder ", Received: 12 March 2023 / Revised: 15 April 2023 / Accepted: 19 April 2023 / Published: 23 April 2023

[19] Nistor, A.; Zadobrischi, E. The Influence of Fake News on Social Media: Analysis and Verification of Web Content during the COVID-19 Pandemic by Advanced Machine Learning Methods and Natural Language Processing. *Sustainability* **2022**, *14*, 10466.

[20] Ali, A.M.; Ghaleb, F.A.; Al-Rimy, B.A.S.; Alsolami, F.J.; Khan, A.I. Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique. *Sensors* **2022**, *22*, 6970

[21] Zhou, M.; Duan, N.; Liu, S.; Shum, H.-Y. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering* **2020**, *6*, 275–290.

[22] Vu, D.T.; Yu, G.; Lee, C.; Kim, J. Text Data Augmentation for the Korean Language. *Appl. Sci.* **2022**, *12*, 3425.