Proceedings of the 9th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC-2025)
DVD Part Number: CFP25OSV-DVD; ISBN: 979-8-3315-7472-7

A Node2vec based Fake News Detection Model using News Source URL

Mosharat Jahan
Dept. of CSE
University of Chittagong
Chittagong-4331, Bangladesh
mosharratruth@gmail.com

Tanjim Mahmud
Dept. of CSE
Rangamati Science and Technology University
Rangamati-4500, Bangladesh
tanjim_cse@yahoo.com

Abubokor Hanip
Washington University of Science and Technology
Alexandria, VA 22314, USA
abu.hanip@wust.edu

Mohammad Shahadat Hossain
Dept. of CSE
University of Chittagong,
Chittagong-4331, Bangladesh
hossain_ms@cu.ac.bd

***Abstract*—The rapid expansion of digital media and the internet has facilitated the spread of fake news, posing serious threats to information integrity and societal harmony. Now is the time of internet where every incident becomes a news and every news is seen all over the world in mere seconds. Which makes the possibility of spreading a false news more prominent. In a country like ours where people tend to be very excited over a simple news, spreading of fake news can be very impactful in many ways. There are numerous examples where a fraud and falsified news has done severe harm to people. So with time many researches are being conducted to detect fake news in different languages. Though most of the researches are conducted regarding high resource languages but the impact of fake news is not confined into language. This thesis presents a novel approach to fake news detection, leveraging web search results and harnessing the power of the node2vec algorithm. The objective of this proposed model is to detect fake news from various Bengali websites as there is not many researches conducted regarding this language.**

***Index Terms*—Bangla fake news, Node2vec, Web search results, NLP**

I. INTRODUCTION

The propagation of false news has grown to be a significant concern in today's quickly expanding digital era, posing substantial dangers to information integrity and public dialogue [1]. It is essential to create strong and effective solutions to address this problem since the ease with which information can be shared on social media and the size of the internet have accelerated the spread of disinformation. Nowadays, news sharing amongst people is similar to an unofficial competition [2]. As absurd as it may sound, it is a truth that individuals now evaluate one another primarily on the number of likes, comments, and views they receive on social media profiles [3]. But some people often use unethical methods to gain access to these beliefs and preferences. The most popular method is to post erroneous or unverified news in order to periodically excite visitors and garner the desired likes, views, and shares. Even while it might not seem harmful at first, not all news has an impact on everyone. News that seems unimportant to us may be of great importance to someone else. People also utilize fake news for their personal gain, which is quite evident when we look at political parties in any nation [4]. Political parties frequently broadcast misleading information about themselves or about their competitor parties in an effort to attract support and win elections. People's faith in online news is now seriously threatened by fake news as well as phoney product or other promotions [5], [6]. Sometimes the news may not be false, but the headline may be embellished to draw attention to something unrelated to the news it contains. The need to identify bogus news on the internet has become unavoidable as consumers these days prefer social media-based news to traditional news printed media. Some websites, such as www.politifact.com, www.factcheck.org, and www.jaachai.com, where the administrators manually update erroneous reports published on the internet with logical and factual justification, can be used to discover false news [5]. However, these websites are unable to operate in real-time and are not equipped to react swiftly to any breaking news. Numerous academics use a variety of techniques to effectively identify bogus news. Researchers and policymakers alike have shown a great deal of interest in fake news identification because the effects of disinformation can result in societal discord, political instability, and a decline in public confidence in the media. In order to find linguistic patterns and textual discrepancies, natural language processing (NLP) techniques are typically used in traditional approaches of fake news identification. While these measures have had encouraging results, they frequently have problems coping with the constantly changing nature of fake news, as offenders constantly modify their evasion tactics. This thesis suggests a novel method for identifying fake news utilizing web search results and the potent node2vec algorithm to address these problems. We can extract rich semantic data from the network of linked web sites using the node2vec algorithm, a well-known node embedding method in graph analytics. This enables us to establish a more

thorough picture of the information ecosystem surrounding a certain news story.

This study's main goal is to create a novel false news detection algorithm that goes beyond existing content-based approaches. We hope to use the collective wisdom of the internet to confirm or deny the veracity of a news piece by combining web search results into the study. This will provide us a distinctive viewpoint and enable us to cross-validate data from other sources, improving the accuracy and resilience of the model.

II. Related work

As stated previously there have many attempts to detect fake news from internet by many researchers. And most of the attempts were made regarding content based features. The news content has many parts like image, text etc but as a huge part is consisted by text so main focus always remains on text based fake news detection. But the limitation in this way is not negligible because even though previously the pattern of text in fake news and real news could be separated, now it is not the case. So text based detection in loosing its credibility with time. To avoid loopholes of text based detection many studies focus on context based features which includes the idea of linguistic features of the news. In this approach user based features and network based features are focused on [7]. Which means the properties of news source and news link are studied to find out if the news is true or not. Just like different features every studies uses different methods for their corresponding work. Most of the previous studies used Support Vector Machine(SVM) [8]. It was intended to show that absurdity, punctuation marks and grammar are best for identifying poignant news by collecting a number of news articles from some news websites.

Links on the internet, a network-based feature, could potentially be a valuable resource for context data. The type of URL and host information are crucial factors in recognizing suspicious URLs, even though they are not related to the identification of fake news [9]. A research examined a number of link attributes to identify bogus news [10]. They created link-related features based on the presumption that connections to legitimate news will have a https prefix, .gov, .co, and.com domain extension. In addition, links to bogus news are sometimes excessively long, brief, complex numerical, or blog hosts. Additionally, they attempted to turn connections into one-hot vectors and utilize them as features for fake news identification, but when used in isolation, they produced no useful results. However, because all the features connected to the linkages were synthesized, they were combined into a single feature variable. By integrating this variable in the model, the baseline model's performance was somewhat enhanced (2.44%), and accuracy of 53.28% was produced. But when compared to other factors like the text variable (64.35%) and the Twitter variable (57.22%), it was discovered that the conversion did not significantly boost performance. Use of https as a feature to assess the credibility of the URL had the opposite effect, with more https -enabled URLs turning up in fake news [11]. Further research is required because their study lacked sufficient links to support it. The two study instances listed above used URL-based source data often for false news detection research, but they were unable to demonstrate a discernible improvement in the performance of the models.

Among all the studies, the research done by Vishwakarma et al. is a noteworthy one. His study proposed URL search based result fake news detection using a set of reliable links. This research shows 85 percent accuracy and which shows the effectiveness of using web search links in detecting fake news. But the idea of the this whole study was not new as it used the concept of white list based phishing detecting [12]. Even after considering the effectiveness of this study there are still some drawbacks of this research which is it needs human participation to collect the set of links.

Another research by Perez Rosa et al. which includes previously mentioned SVM model used only linguistic features of text. They also collected a dataset of 240 legitimate news from different news websites in US and made another dataset consisting the fake version of news from the previous dataset. They used crowd sourcing to generate this fake version of news using Amazon Mechanical Turk(AMT) [13].Moreover as we can see here also human intervention is needed to collect the relevant dataset and also to generate false version of the news collected in the dataset. Moreover to extract more secretive characteristics of false news neural network based model are undoubtedly more reliable.

There are also some researches of clickbait detection which also uses linguistic features and neural network with dataset that normally contains some click baits from various news websites [14]. Not only English but there are some studies done on a couple of low resource languages like Indonesian using Naive Bayes classifier [15].

III. Methodology

This section outlines the research approach used to create the node2vec-based fake news detection model for the Bangla language using web search results.

A. Data Collection

The data gathering procedure for creating the fake news detection dataset in Bangla is described in this section (see table I). To ensure the collecting of trustworthy and varied news pieces, web scraping methods and the choice of dependable news sources are described. To prepare the text data for upcoming comparison model training and evaluation, the preprocessing steps, which include tokenization, stop-word removal, and stemming [16]–[20], are described in depth. The information should fit the criteria of false news because it has the story's content. 'Banfakenews Dataset' is an experimental dataset that we created for this study [5].

We obtained 242 different categories from our dataset since various outlets categorize the news in different ways. We combined related categories from several news sources to make a single category in order to generalize it. Finally, we

TABLE I: Dataset Details Information

Type of Article	Total Data	Percentage
Real	14,707	81.83%
Fake	3,264	18.16%

TABLE II: Number of News in Each Category

Data Type	Fake	Real	Total	Fake %	Real %
Crime	72	310	382	18.8	81.1
Education	55	330	385	14.2	85.7
Entertainment	182	783	965	18.8	81.1
Finance	2	378	380	0.52	99.4
International	154	2166	2320	6.63	93.3
Lifestyle	183	280	463	39.52	60.4
National	168	5610	5778	2.90	97.0
Politics	166	886	1052	15.77	84.2
Sports	99	1959	2058	4.81	95.1
Technology	48	259	307	15.63	84.3

divide the dataset's news stories into the 12 categories (Table 2).

B. Data Preprocessing

In this step each title of fake and real news from the collection is now typed as a search query into an online search engine. The first links from each search result are also gathered when a lot of links are indexed as a consequence of a web search(see figure 1). In this study, Google was the web search engine we used to put our suggested research model into practice. A ranking mechanism called PageRank was created by Larry Page and is used in part to determine the order of search results returned by Google. According to Google2, PageRank calculates a rough evaluation of a website's significance by counting the quantity and calibre of links pointing to a page. The essential premise is that more significant websites will probably get more links from less important websites. Google's search is made to prioritize websites with high PageRank content and high relevance to search phrases at the top of the search results page through a number of technologies, including PageRank. After collecting the links, preprocessing these links is applied. Unnecessary sub-paths from the links are deleted, and the domains are extracted and identified. Consequently, each link is collected and stored in the dataset.

C. Applied Algorithm

Graph Representation: The node2vec algorithm receives a graph as input that is represented as either an adjacency list or an adjacency matrix. This graph representation, which serves as the foundation for the random walk process, records the relationships between nodes(see figure 2).

Biased Random Walks: Node2Vec conducts biased random walks to investigate the topology of the graph and provide node sequences for each walk. Biased random walks establish a balance between discovering more extensive graph topologies and exploring nearby neighbourhoods. The chances of the walker returning to the node they came from and the likelihood of exploring nodes in the same neighbourhood or in different regions of the graph, respectively, are controlled

Fig. 1: Link collection process according to news title.

by the two parameters return and in-out, which are used to do this.

Word2Vec Adaptation: In the context of the Word2Vec method, biased random walk-generated node sequences are viewed as sentences. Assuming that each node in the graph represents a word, Word2Vec is used to extract embeddings from these sequences(see figure 3). The goal of Word2Vec is to guarantee that co-occurring nodes in these random walks have similar embeddings.

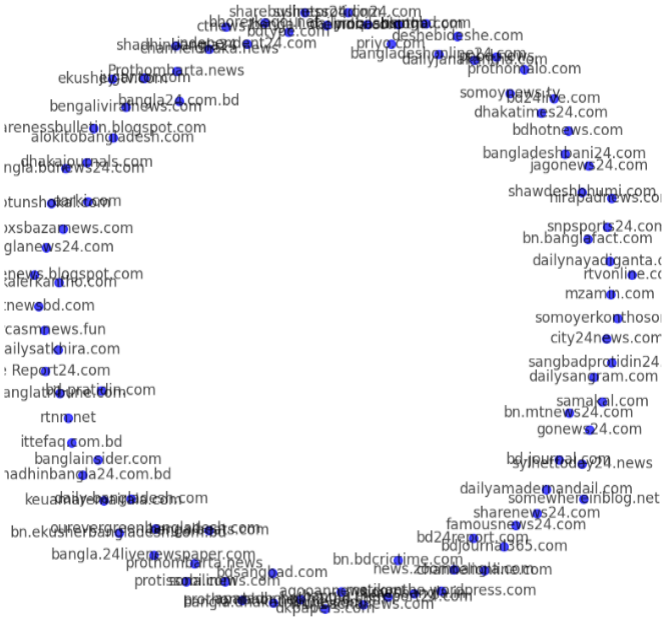

Fig 2: Web Graph for Node2vec Model

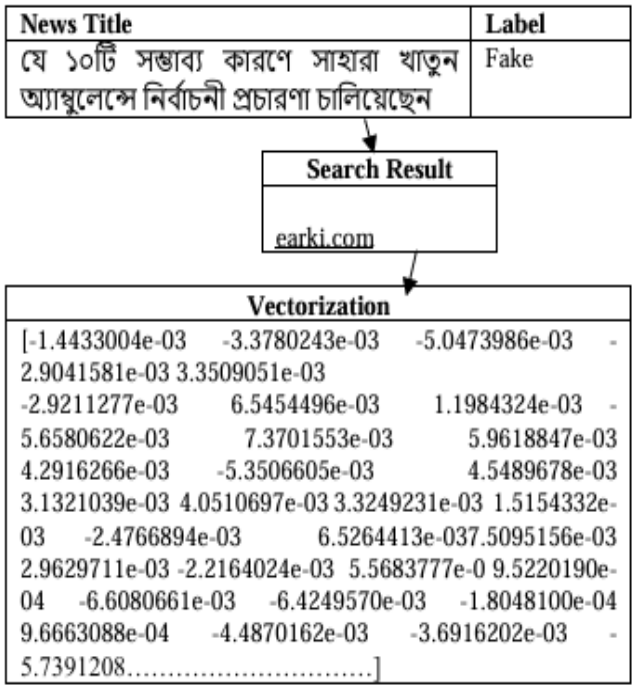

Fig 3: Link Vectorization for Feature Extraction

Embedding Generation: Low-dimensional embeddings are created for each node in the graph after the Word2Vec conversion is finished. These embeddings are dense vectors that generally have hundreds of dimensions and exist in a continuous vector space. They reflect each node's connections to other nodes in the graph and its structural and semantic features(see figure 4).

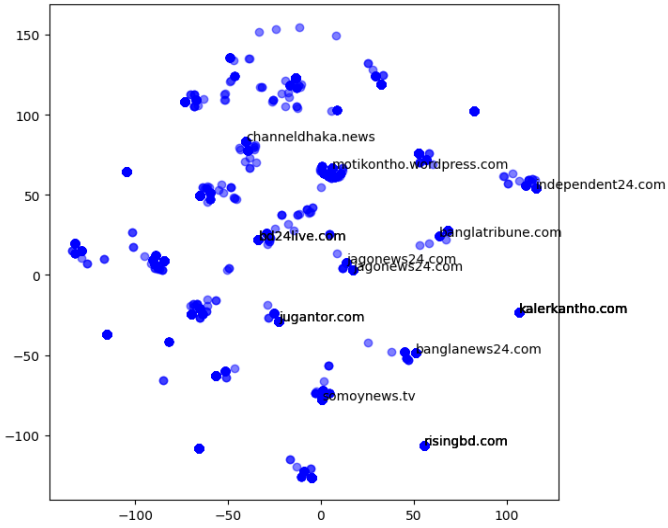

Fig 4: Embedded Nodes in Graph

IV. Results

In this section, we give a tabular comparison of the accuracy results obtained by various models when it comes to the identification of fake news.

TABLE III: Results for SVM Model

Metric	Proposed Model	Comparison Model
Test Accuracy	0.99	0.96
Train Accuracy	0.99	0.98
Precision	0.99	0.97
Recall	1.0	0.98
F1 Score	0.99	0.98

TABLE IV: Results for CNN Model

Metric	Proposed Model	Comparison Model
Test Accuracy	0.87	0.86
Train Accuracy	0.86	0.86
Precision	0.87	0.86
Recall	1.0	1.0
F1 Score	0.93	0.92

A. Discussion

The experimental findings from comparing several false news detection models employing Node2Vec embeddings and text-based representations are thoroughly discussed in this section. In the broader context of false news identification, we examine the consequences of each model's performance analysis. With regard to both Node2Vec and text-based representations, the Support Vector Machine (SVM) model performed quite well. The model's ability to accurately discern between instances of true and fake news is demonstrated by the astonishingly high accuracy scores, which exceed 99%. The SVM model's accuracy, recall, and F1-score were all outstanding, demonstrating how well it could categorize a large number of legitimate news articles and identify every case of fake news(see table III and figure 2)). For both Node2Vec and text-based embeddings, the Convolutional Neural Network (CNN) model displayed competitive accuracy and performance. The CNN model's accuracy of 86%, but significantly lower than SVM, shows its effectiveness in distinguishing between authentic and false news articles(see table IV and figure 2, 3). Additionally noteworthy were the model's precision, recall, and F1-score, which showed how well it could identify intricate patterns in textual material.

For both Node2Vec and text-based embeddings, the accuracy of the Logistic Regression model was commendable. The model is effective at identifying bogus news, as evidenced by accuracy scores that approach 97%(see table V, figure 4 and 5). Its high precision, recall, and F1-score further demonstrated its propensity for making accurate predictions. When we compare the models' performances, we find that the SVM model had the highest accuracy of all the methods. With 99% or higher F1-scores, precision, recall, and robustness in detecting bogus news, it proved to be effective. Although slightly less effective

TABLE V: Results for Logistic Regression Model

Metric	Proposed Model	Comparison Model
Test Accuracy	0.97	0.95
Train Accuracy	0.97	0.95
Precision	0.97	0.88
Recall	1.0	0.99
F1 Score	0.98	0.97

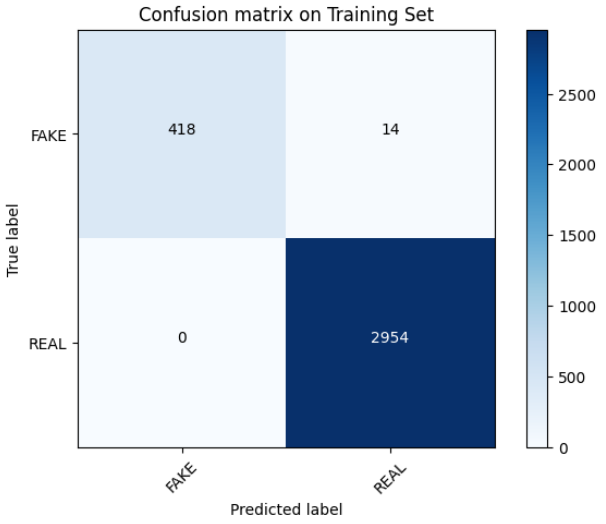

Fig. 2: Confusion Matrix for SVM (Node2vec)

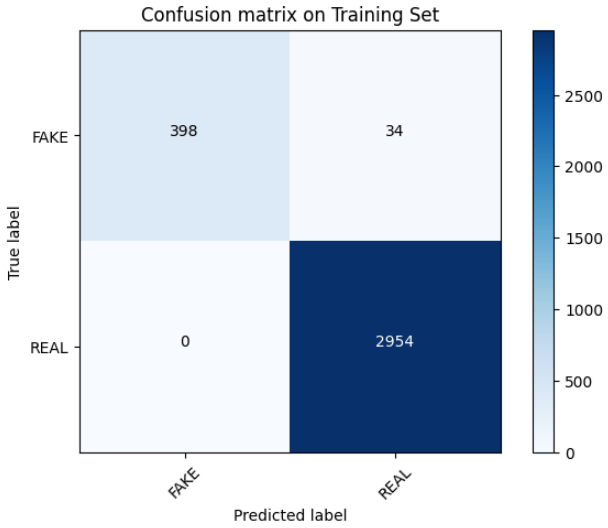

Fig. 4: Confusion Matrix for Logistic Regression (Node2vec)

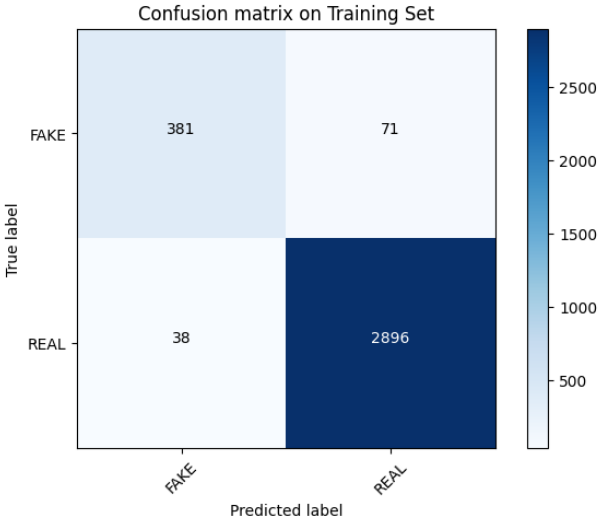

Fig. 3: Confusion Matrix for SVM (Text-based)

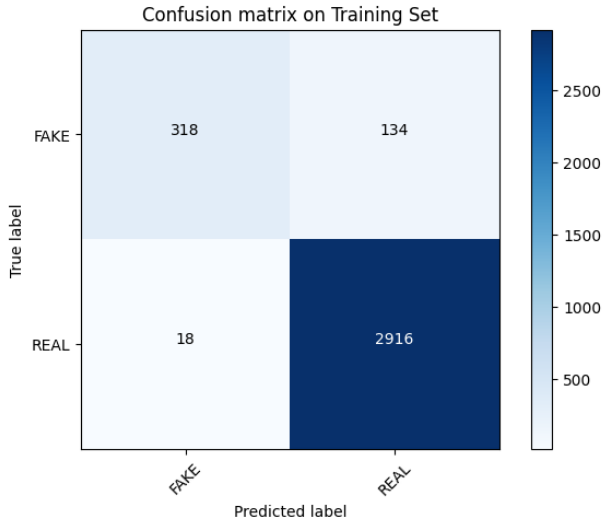

Fig. 5: Confusion Matrix for Logistic Regression (Text-based)

than SVM, the CNN model nonetheless performed admirably, with an accuracy of more than 86%.

In terms of accuracy and false positive rates, our suggested Node2Vec-based model fared better than the text-based model. The Node2Vec model produced predictions with a better degree of accuracy, demonstrating its efficacy in identifying significant linkages in the web graph. The text-based model, in contrast, showed a greater false positive rate, indicating its limits in identifying instances of actual and fake news. Our suggested Node2Vec-based model clearly outperforms text-based models when performance is compared. In comparison to the text-based model, it delivers higher accuracy and reduced false positive rates. By utilizing web graph embeddings, the Node2Vec model is able to capture significant relationships between news sources, producing predictions that are more precise. The text-based model, on the other hand, is purely dependent on text and may have trouble accurately differentiating between true and false news.

V. Conclusion

We summarize our research on identifying false news using text-based representations and Node2Vec-based embeddings in this section along with our main conclusions. We go over the benefits of our suggested model over the comparison model, as well as how well it can manage datasets that are unbalanced, and we also point out some of the Node2Vec algorithm's and our suggested approach's shortcomings. The suggested Node2Vec-based fake news detection technique outperforms the conventional text-based comparison approach by a wide margin. First off, the use of graph embeddings considerably reduces the execution time. Our approach is ideally suited for real-time applications and large-scale datasets because of its

improved efficiency, which enables prompt and precise predictions. Second, our model demonstrates robustness in managing imbalanced datasets, a common problem in the identification of fake news. By utilizing Node2Vec embeddings, it effectively captures subtle patterns and significant relationships inside the online network, resulting in precise predictions even in the presence of a small number of fake news occurrences. Furthermore, by utilizing Node2Vec-based embeddings, the model is able to extract contextual data from the online graph, assisting in the identification of bogus news websites and creating links between them. The web graph representations help the machine recognize trends and classify bogus news sources, enabling more precise categorization. Node2Vec has certain drawbacks even if it offers useful embeddings for our fake news detection model. Its sensitivity to the selection of hyperparameters, such as the walk length and the number of walks per node, is a severe constraint. To acquire the best outcomes, these factors must be fine-tuned over a long period of experimentation. The accuracy and comprehensiveness of the web graph data may differ depending on the sources' accessibility, which can also have an impact on the algorithm's performance. Although Google dominates the web search market, it is preferable to check node2vec's usability using additional search engines, like Bing and Yahoo. We intend to use a variety of search engines to further test and validate our fake news detection technology in upcoming studies. Additionally, our suggested model has some restrictions. Its dependency on the availability of online graph data is on of its limitations. The model's performance might be impacted in situations when access to web sources is constrained. The quality and amount of the dataset also affect the performance of our suggested method, as is the case with any machine learning model. The accuracy and generalization abilities of the model could be further improved with a larger and more varied dataset.

REFERENCES

[1] N. Absar, T. Mahmud, A. Hanip, and M. S. Hossain, "Semi-supervised based bangla fake review detection: A comparative analysis," in *2025 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2025, pp. 1428–1433.

[2] S. U. Habiba, F. T. Johora, T. Mahmud, T. Tasnim, F. Tasnim, S. Kabir, M. S. Hossain, and K. Andersson, "Enhancing low-resource bangla fake news detection through deep convolutional neural networks," in *International Conference on Intelligent Computing & Optimization*. Springer, 2023, pp. 104–114.

[3] S. U. Habiba, T. Mahmud, S. R. Naher, M. T. Aziz, T. Rahman, N. Datta, M. S. Hossain, K. Andersson, and M. S. Kaiser, "Deep learning solutions for detecting bangla fake news: A cnn-based approach," *Proceedings of Trends in Electronics and Health Informatics: TEHI 2023*, p. 107.

[4] T. Mahmud, T. Akter, M. T. Aziz, M. K. Uddin, M. S. Hossain, and K. Andersson, "Integration of nlp and deep learning for automated fake news detection," in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*. IEEE Computer Society, 2024, pp. 398–404.

[5] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, "Banfakenews: A dataset for detecting fake news in bangla," *arXiv preprint arXiv:2004.08789*, 2020.

[6] T. Mahmud, I. Hasan, M. T. Aziz, T. Rahman, M. S. Hossain, and K. Andersson, "Enhanced fake news detection through the fusion of deep learning and repeat vector representations," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2024, pp. 654–660.

[7] J.-S. Shim, Y. Lee, and H. Ahn, "A link2vec-based fake news detection model using web search results," *Expert Systems with Applications*, vol. 184, p. 115491, 2021.

[8] V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the second workshop on computational approaches to deception detection*, 2016, pp. 7–17.

[9] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: learning to detect malicious web sites from suspicious urls," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 1245–1254.

[10] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, and P. Nakov, "Predicting factuality of reporting and bias of news media sources," *arXiv preprint arXiv:1810.01765*, 2018.

[11] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, "Credibility-based fake news detection," in *Disinformation, misinformation, and fake news in social media: Emerging research challenges and Opportunities*. Springer, 2020, pp. 163–182.

[12] L. Li, E. Berki, M. Helenius, and S. Ovaska, "Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate?" *Behaviour & Information Technology*, vol. 33, no. 11, pp. 1136–1147, 2014.

[13] V. L. Rubin, "Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions," *Journal of documentation*, vol. 75, no. 5, pp. 1013–1034, 2019.

[14] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, 2016, pp. 9–16.

[15] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of hoax news detection using naïve bayes classifier in indonesian language," in *2017 11th International Conference on Information & Communication Technology and System (ICTS)*. IEEE, 2017, pp. 73–78.

[16] T. R. Das, T. Mahmud, A. Hanip, and M. S. Hossain, "Harmful tweet detection using supervised and semi-supervised learning techniques," in *2025 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2025, pp. 1421–1427.

[17] R. Tasnia, T. Mahmud, A. Hanip, and M. S. Hossain, "Leveraging semi-supervised learning and generative adversarial networks with transformer and flair embeddings for detecting patronizing and condescending language," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2025, pp. 1–5.

[18] A. Majumder, T. Mahmud, T. Barua, N. Jannat, M. F. B. A. Aziz, D. Islam, R. Chakma, M. S. Hossain, and K. Andersson, "Harnessing bert for advanced email filtering in cybersecurity," in *2025 8th International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2025, pp. 66–71.

[19] T. Mahmud, G. S. Hossain, M. H. Ali, T. Hasan, M. F. B. A. Aziz, M. T. Aziz, M. S. Hossain, and K. Andersson, "A machine learning-based framework for malicious url detection in cybersecurity," in *2025 8th International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2025, pp. 61–65.

[20] T. Mahmud, M. Ptaszynski, and F. Masui, "Leveraging explainable ai and sarcasm features for improved cyberbullying detection in multilingual settings," in *2024 IEEE Digital Platforms and Societal Harms (DPSH)*. IEEE, 2024, pp. 1–8.