Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Detect Misinformation Using Two Stage Semantic Extractor Based On Neural Network Classification

Roshan R.Karwa[1] and Sunil R Gupta[2]

[1]Department of Computer Science & Engineering ,PRMIT&R, Badnera, Maharashtra, India.

[2] Department of Computer Science & Engineering, PRMIT&R, Badnera, Maharashtra India.

ABSTRACT

The tremendous use of social media is the one of the important cause of generation of huge quantity of data. Analyzing this huge data is very important to get insights from the data and apply to solve real life problems. There is no accurate medium to check the semantics and authentication of data being generated. Any user of social media can post whatever they think according to their own perspection and opinion as well as user share the information without checking the authenticity, that is impacting societ in various ways. Many researchers are using Artificial Intelligence based algorithms which gives idea of detecting misinformation(commonly refereed as Fake News) potentially. Many of this techniques rely on the dataset being chosen to solve the problem. They mostly designed based on the direct feature. Understanding context with respect to its semantic is very necessary. Thus to overcome, this paper introduces Two Stage Semantic Extractor based Neural Network Method (TSENNM). According to experimental results, the proposed model obtained an good accuracy when compared with the previous model.

Keywords: Artifical Intelligence, Deep Learning, Fake news, Misinformation, Neural Network. Semantic Feature extraction,

1. INTRODUCTION

Today every mobile consumer is using social media due to its easy accessibility and less cost. Every real time application is connected or appended with social media [1]. As social media become core part of everybody’s life, people used to share their thoughts, ideas, opinion, daily activities on social media. With this sharing of own things, people used to share forwarded information as well. Journalist or news channels are also posting the current affairs over social media with traditional media and due to the user friendly usage of social media, people are prefering social media to get daily affairs/information instead of traditional sources such as newspapers or television. The main source of this information is user. People read the information and believe it without verifying it and if it is false information then it impact social decisions, social environment [2]. Social media consists of Facebook, Twitter, Snapchat, Instagram, Whatsapp etc. In the western region, twitter is often considered as platform where people get across false news, whereas in rural region whatsapp and facebook is setting the benchmark in spreading the information. Today the facebook and whatsapp application company is also trying to build application to combat the misinformation [3]. Impact on Financial aspects and people perspective are main motive to spreading misinformation [4]. Mis information is not only limited to Political, Spiritual, Financial etc domains; but also changing perspective of public from bad to good or good to bad [5]. Stock market is also affected by the false news, it change the sentiment of investor for their valuability in investing shares [6,7]. During election it is always observed that if some news spread about any candidate or party then it change the perspective of common people about the party or candidate [8,9].

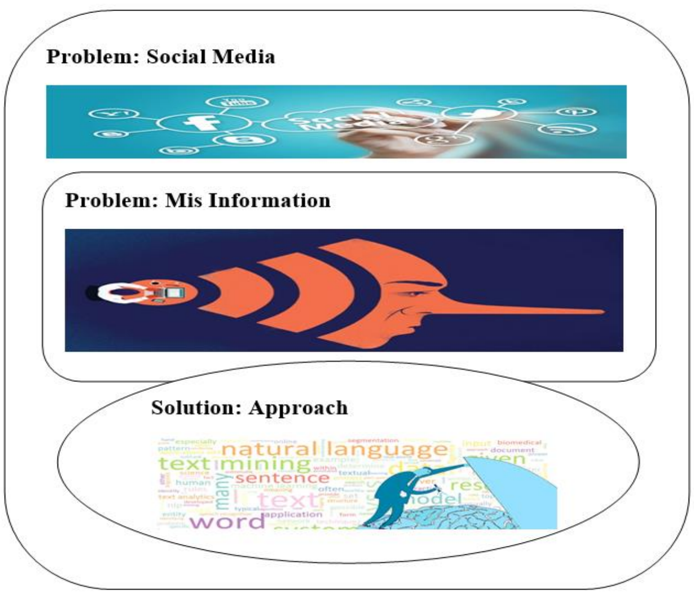

Figure 1: Problem and Solution of using Social Media

Manual validating authenticity of information (news) by human being is very difficult, biased and time consuming. Today researchers have been developing and designing techniques with the enhancement of Artificial Intelligence to know, identify, and reduction of false information. This technique includes Machine Learning, Natural Language Processing and Deep learning with getting insights of data with domain Data Science.

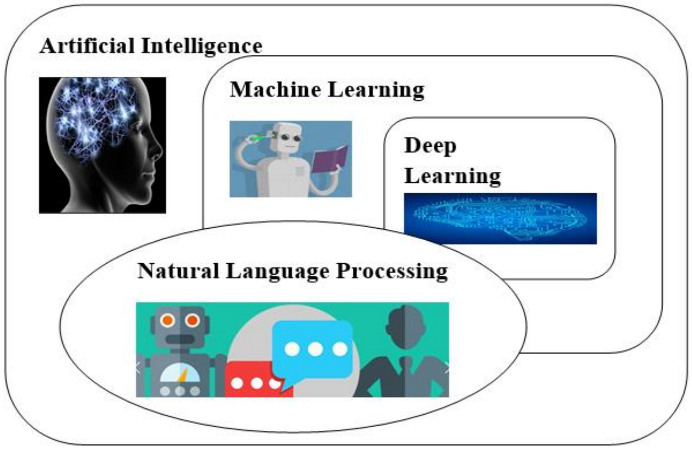

Figure 2: Artificial Intelligence Techniques to solve problem of misinformation

Researchers target is to identify and recognize false information automatically. The automatic authentication of a text item as genuine or false is a difficult problem [10]. Distinguishing real and false news i.e. information needs improvement as per looking in the current research direction. Conseuently, system should be develop and design to detect misinformation so that it can be stop or reduce through the use of artificial intelligence techniques [11,12]. As a result, our research proposes Two Step Semantic Neural Network Method. In this proposed model, there are two main layer. The first layer is of obtaining semantic features and second layer is predicting the news(piece of information) with the input of feature vector.

The residual preparation of research paper is as follows: Background work is present in Section 2. Section 3 consists proposed methodology. Implementation and Result is presented in Section 4 whereas Section 5 concludes the paper.

2. LITERTURE REVIEW

Misinformation is propogated on social media soon. It result into the chaos in the society in terms of finance and perspections. Therefore, it is very necessary to combat the misinformation. System which is already presented and developed by different researchers are discussed in this section.

Karimi et al [13] recommended the MMFD framework, that integrates automated feature engineering, multi-source merging, as well as automated degrees of fakeness identification into a comprehensible and understandable mode, in which CNN examines spatial relationships of each text in a claim and LSTM identifies temporal dynamics in the claim. This model incorporates the best characteristics of both models, because LSTM performs better for longer words.

The problem formulations, datasets, and NLP solutions generated for this task were thoroughly studied and compared by Oshikawa et al [14], who also acknowledged their potentials and limitations. The researchers also stressed the distinction between misleading news identification and other related problems, as well as the importance of natural language processing (NLP) solutions for detecting misinformation.

Zhang et al. [15] suggested a novel analytics-based strategy for false text detection. Paper starts by discussing the architecture of the proposed technique, and also the core mathematical framework, including implementation details and assessment using a corpora of data sets. The researcher gathers legitimate and malicious news, which is then transformed from a document-based corpus into a topic and event-based representation. A two-tiered technique for classification and prediction of false information is used, which includes recognising phoney themes and fake events. The efficacy of the proposed technique is demonstrated by the development and validation of a breakthrough Fake News Detection system.

Leveraging machine learning technologies and natural language processing, Smitha et al [16] developed a system and method for recognizing inaccurate information in media articles. Numerous variable selection techniques, including count vector, TF-IDF, and word embedding, are used to construct the feature map in this research design. To figure out whether the news is fake or real, seven different Machine Learning Classification algorithms were used.

Ozbay et al [17] presented a two level strategy to detect misinformation. To convert unorganized data sources into organizing data, the first stage of the approach comprises employing a range of preprocessing techniques to the data collection. Vectors produced with the acquired TF weighting technique and Document-Term Matrix encapsulate the words in the news data set. In the second stage, information retrieval methodologies were used to apply artificial intelligence algorithms to the organized data.

Lu et al [18] created Graph-aware Co Attention Networks (GCAN) model which is a novel neural network-based model . Furthermore, it anticipates misinformation essence of the original post and the persons who spread it. Moreover, GCAN has a good track record of timely detection of false information. GCAN, according to the researchers, may be used for more than simply detecting fake news on social media; it can also be used for emotion classification, offensive speech identification, and twitter post popularity forecasting.

Ibrishimova et al [19] discussed false propaganda in the regarding data warfare and found that the most viable approach for detecting misinformation combines source and fact verification with NLP assessment, and the researcher suggested a combination of framework based on previous task in automating event categorization.

Kaur et al. [20] created a novel multi-level ensemble voting mechanism. Three feature extraction procedures are used to feed the corpus's collection of characteristics into the specified machine learning models.

Kaliyar et al [21] proposed combining of many concurrent blocks of a single-layer deep Convolutional Neural Network (CNN) with varying kernel sizes and filters with the BERT, a deep learning technique (FakeBERT) based on BERT (Bidirectional Encoder Representations from Transformers). This combination excels at coping with ambiguity, the most challenging challenge in natural language processing.

Verma et al [22] explored number of perspectives on how to identify such false news online and proposed the P2C2 methodology. The method is divided into two parts: detection and verification. In the detection stage, the credibility of the news source and patterns of the news are considered. In the verification phase, the relevance of the information is assessed.

Nasir et al. [23] focused on challenge given by social media and introduced Artificial Intelligence based technique to detect mis information. The research introduced hybrid deep learning model consisting both Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) both which improves the performance of the proposed false news detection model.

Table 1: Recent Advance and research gaps presented by researchers

Author	Technique	Research Challenge
Karimi et al [13]	MMFD with LSTM	More sources, such as temporal data, social networks, and user interactions, should be included.
Oshikawa et al [14]	NLP based technique	Examine if hand-crafted features may be coupled with neural network models, as well as the suitable use of non-textual data and the extension of the verification method to include contents.
Zhang et al. [15]	Data Analytice based strategy	To process real-time attributes of news stories, include a pre-processing module.
Smitha et al [16]	Feature Vector strategy	For greater accuracy, a dataset with a higher number of articles from various sources could be employed, as it contains more jargon and significant content.
Ozbay et al [17]	Two level strategy-Preprocessing & Detection	Integrating intelligent optimization algorithms for better

Lu et al [18]	Graph-aware Co Attention Networks (GCAN)	To improve efficiency and explainability, remove event-specific features.
Ibrishimova et al [19]	Hybrid ML with NLP	Consider semantic similarity
Kaur et al. [20]	Multi-level ensemble	No annotated dataset available for images
Kaliyar et al [21]	FakeBERT	integrating information from news items on a content, context, and chronological level
Verma et al [22]	P2C2 methodology (Propagation, Pattern, Comprehension & Credibility)	Extract Deep Semantic Features
Nasir et al. [23]	Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)	More complex neural network model to be considered

As discussed, researchers have been developing techniques to combat fake news i.e. to prevent(reduce) spread of misinformation using Artificial Intelliegence. Still accuracy of detection and prevention presents many challenges to solve. Our research create a new artificial intelligence system to detect mis information in social media to solve above mentioned problems.

3. TWO STAGE SEMANTIC BASED NEURAL NETWORK MODEL

The development of the World Wide Web and the widespread acceptance of social networking sites (such as Facebook and Twitter) laid the foundation for extraordinary process of knowledge sharing in mankind's life, however classifying a written content as misleading or disinformation is difficult to automate. To address this, our research presents a Two level Semantic based Neural Network Model, which uses Semantic Structure Model and Deep Convolutional Neural Network models to identify and classify false news, as a deep learning technique for automated news article categorization. The input strings from the provided input data are fed into the input layer, where subsequent convolutional layers extract local contextual information and a max-pooling layer produces a global feature vector. Figure 3 depicts the suggested architecture.

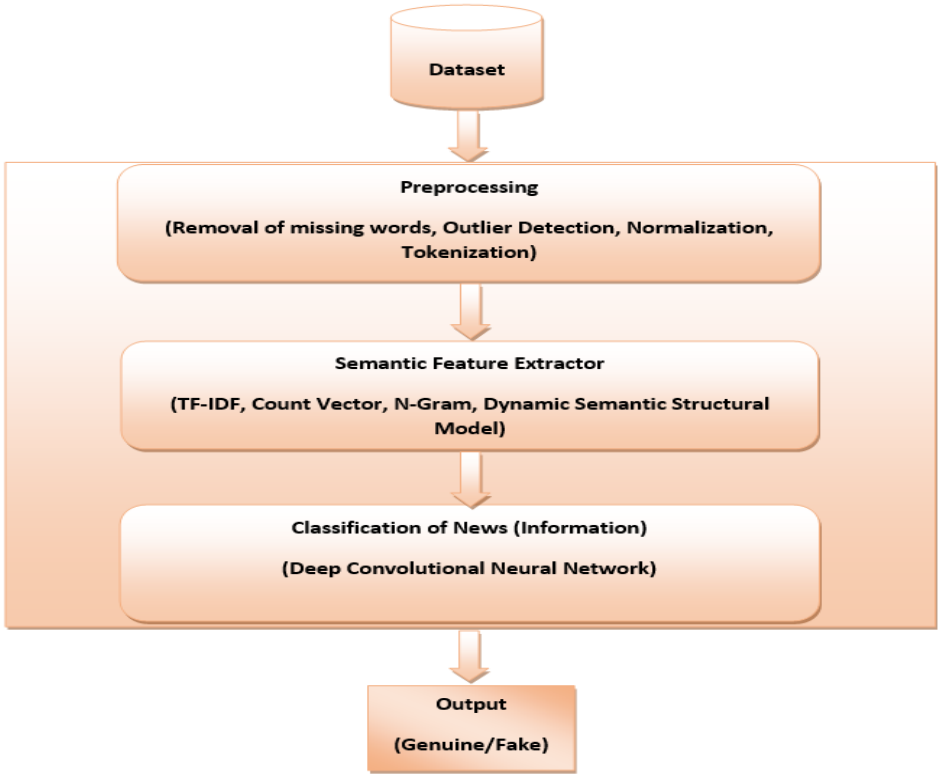

Figure 3: Two Level Semantic Neural Network Model

3.1 Semantic Extractor Model

Excessive use of social media does indeed have consequences for society, civilization, and industry, with both benefits and risks. Able to detect false propaganda is a challenging task because it requires models to summarise the news and compare it to genuine news in order to determine whether it is fake. Using semantic extractor model and an upgraded neural network model based on the Twitter dataset, the proposed approach detects and classifies bogus news. The input layer in this model is fed with strings from the Twitter dataset. A word bag is constructed from these strings. The preprocessed bag of words is then fed into the Semantic Extractor model, where convolution is a sliding window-based feature extraction approach that captures the context of a word

3.2 Deep Neural Network Model for Classofication

Output of first level is fed into deep convolutional neural networks, which are then used to do classification. A couple of fully connected layers and a max-pooling layer are carried out in triplicate in the proposed Deep Neural Network Model design, followed by a flattening layer and two fully linked layers. All fully connected layers have a kernel size of 3*3 with a stride of 1, whereas all max-pooling layers have a kernel size of 2*2 with a stride of 1. ReLu is used to activate hidden layer functions. Softmax is utilised as an activation function since the output layer conveys the phoney news.

As a result, the likelihood of a class being given to the assertions increases. The binary output is the end result. The output layer returns either 0 (fake news predicted yes) or 1 (fake news predicted no). As a result, the suggested hybrid model performs better in detecting and characterizing fake news in social media.

4. IMPLEMENTATION AND ANALYSIS

The findings of the implementation, and the efficiency of the developed framework, are described in this section. In addition, the baseline approach's comparison results are discussed.

4.1 Dataset Description

LIAR [24] is a publicly accessible dataset for classification and prediction of fake news. POLITIFACT.COM, that encompasses analytical reports including references to source materials for every instance, gathered 12.8K hand labelled brief remarks in diverse situations over the course of a decade. This resource could also be used for fact-checking studies. This new dataset is an order of magnitude larger than prior publicly available false news datasets of a comparable nature. The LIAR dataset contains 12.8K human-labeled short statements from the POLITIFACT.COM API, with each statement being checked for truthfulness by a POLITIFACT.COM editor. Figure 4 depicts sample data.

0	2635.json	false	Says the Annies List political group supports ...	abortion	dwayne-bohac	State representative	Texas	republican	0.0	1.0	0.0	0.0	0.0	a mailer
1	10540.json	half-true	When did the decline of coal start? It started...	energy,history,job-accomplishments	scott-surovell	State delegate	Virginia	democrat	0.0	0.0	1.0	1.0	0.0	a floor speech.
2	324.json	mostly-true	Hillary Clinton agrees with John McCain "by vo...	foreign-policy	barack-obama	President	Illinois	democrat	70.0	71.0	160.0	163.0	9.0	Denver
3	1123.json	false	Health care reform legislation is likely to ma...	health-care	blog-posting	NaN	NaN	none	7.0	19.0	3.0	5.0	44.0	a news release
4	9028.json	half-true	The economic turnaround started at the end of ...	economy,jobs	charlie-crist	NaN	Florida	democrat	15.0	9.0	20.0	19.0	2.0	an interview on CNN

Figure 4: Sample of Dataset

4.2 Sentence Tokenization

After general preprocessing steps such as removal of missing words, outlier detection; tokenization is performed. Sentence tokenization is the method of breaking a text into distinct sentences. The objectionable content is preserved in replacement form during the tokenization process because this sentence tokenizer does not separate individual words. Following the production of individual sentences, reverse substitutions are carried out, resulting in a group of upgraded sentences that closely resemble the original sentence. The length of a phrase is obtained at the preprocessing stage in the form of a histogram, as shown in figure 5.

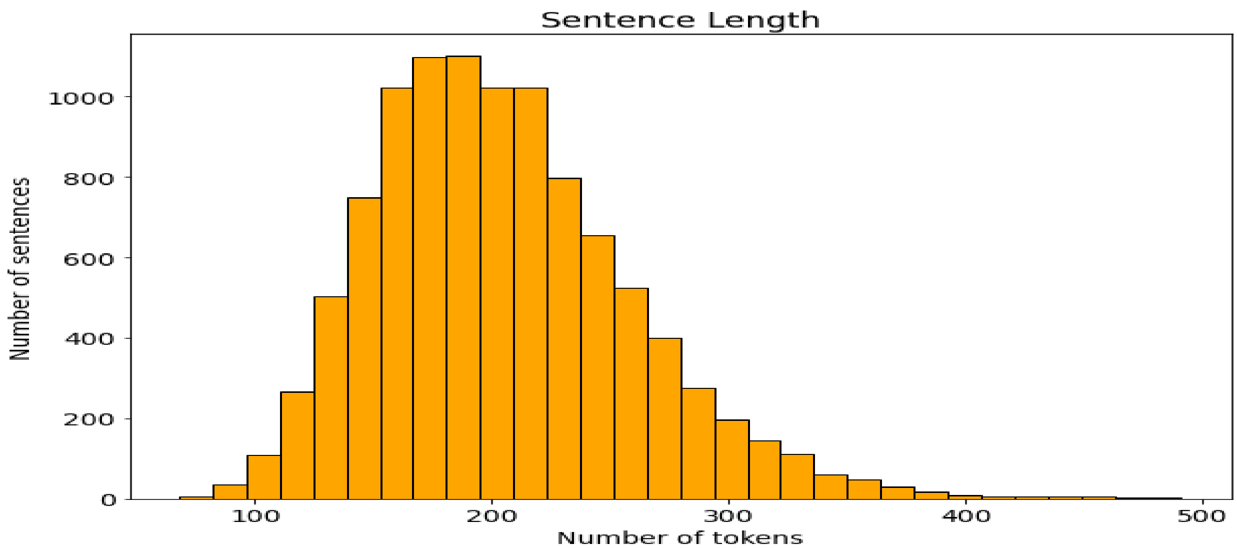

Figure 5: Sentence Tokenization

4.3 Retrieval of Features

Semantic structural model is used in retrieval of features. First, the text is divided into n-grams (n=1,2,3) like unigrams, bigrams, and trigrams.

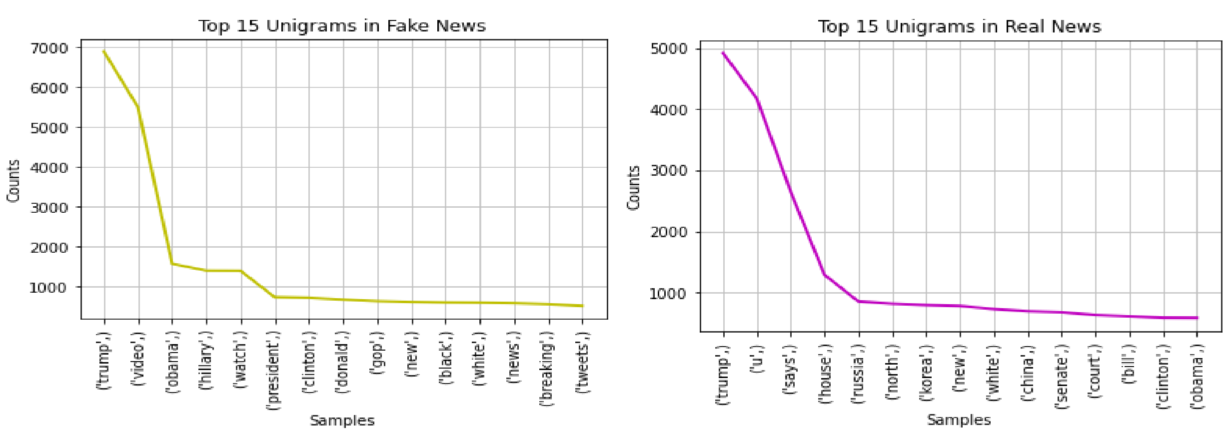

Figure 6: Sample Unigram Feature Extractor

Figures 6 illustrates the topmost 15 most typically used unigrams in fake and real news from the dataset. Both utilize similar terminology like 'trump.' However there are some significant variances. For example, fake news tells that John McCain has done nothing to assist veterans which are said by Donald Trump, but real news claim that Donald Trump opposes marriage equality. He wishes to return.

In the same way, bigram and trigram is extracted. Initially, hashing is applied at the start and the end marks of the sentence in the dataset. Then the word is broken down into bigrams and the semantic feature is extracted by semantic structural model

Furthermore, according to this research, incrementing the n-gram size degrades the performance when classifiers are used. With more extracted features, our research reduces errors. Semantic Extractor performs better in extracting features because both false and real news evaluations featured similar terms. Fake news uses more filler/functions and content keywords than legitimate text. According to our observations, they also use more verbs and adverbs than actual news. True news, on the other hand, employs a greater number of nouns and adjectives.

4.4 Categorization of Real and Fake News

Predicted values are labeled Positive and Negative in our research, whereas actual values are labeled Real and Fake. In figure 10 True positive value is 4890, the True Negative value is 5785, the False Positive value is 435, and the False Negative value is 429. This confusion matrix is essential for determining performance metrics like recall, precision, specificity, and accuracy.

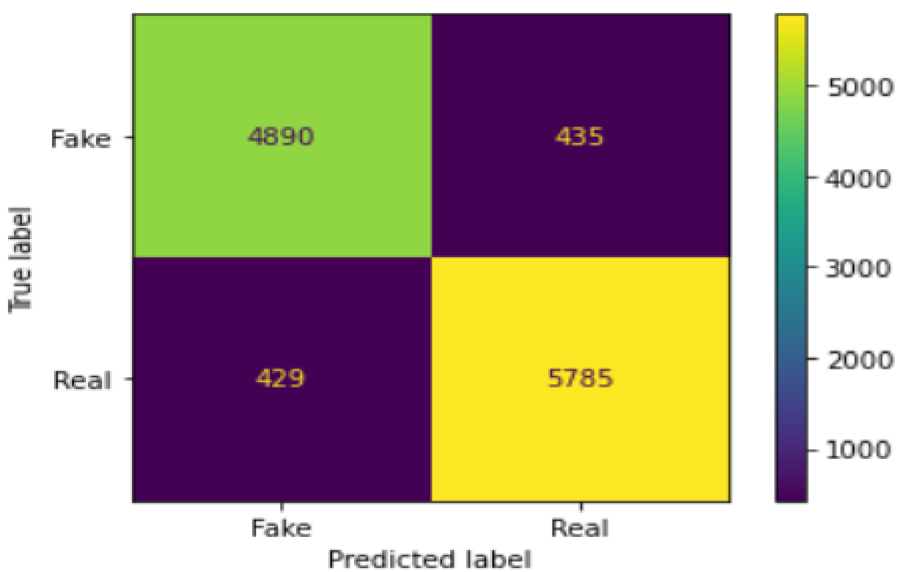

Figure 7: Categorization of Real and Fake News

In investigation, predicted values were designated Positive and Negative, whereas actual values were labeled Real and Fake. True positive value is 4890, True Negative value is 5785, False Positive value is 435, and False Negative value is 429 in figure 7. This confusion matrix is necessary for calculating performance metrics such as recall, precision, specificity, and accuracy.

Following evaluation parameters such as accuracy, F1 Score, precision, and recall are used to assess the proposed technique's effectiveness in detecting and classifying fake news.

Figure 8 illustrates the performance evaluation metrics of the proposed method. The obtained values of accuracy, F1 Score, Precision, and Recall is 92.50%, 92.46%, 92.47%, and 92.45%. The performance of our proposed attains higher accuracy, F1 score, precision, and recall by utilizing two stage semantic extractor and deep neural network based model

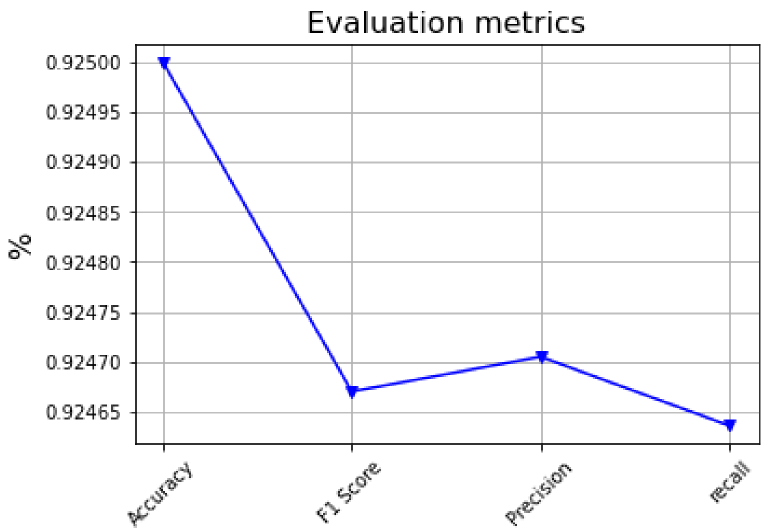

Figure 8: Evaluation Metrics of Proposed System

5. CONCLUSION

Many consumers prefer to get their news and information from social media rather than conventional sources. Fake news has been spread via social media, having negative impacts on people and society in view of finance and ideology both. This paper provides novel Semantic Extractor and improved deep neural network for combat fake news. The results show that the our research is better than the other classifiers in terms of accuracy. The projected findings of this study are remarkable, with a 92.60 percent accuracy rate when using the LIAR dataset, which is exceptionally good at detecting bogus news.As a result, the suggested study is incredibly useful for detecting misleading news (misinformation) and improving accuracy. Furthermore in future, the work can be more enhanced by considering multimedia dataset as well.

REFERENCES

1. Y. Liu and T. Bakici, "Enterprise social media usage: The motives and the moderating role of public social media experience," Computers in Human Behavior, vol. 101, pp. 163–172, 2019, doi:10.1016/j.chb.2019.07.029.

2. M. Pivec and A. Maček, “Employment background influence on social media usage in the field of European project management and communication,” Journal of Business Research, vol. 94, no. pp. 280–289, 2019, doi: 10.1016/j.jbusres.2018.03.021.
3. SoroushVosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. Science 359, 6380 (2018), 1146–1151.
4. Hunt Allcott, Matthew Gentzkow, Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives **31**(2), 211–236 (2017).
5. Nitin Jindal, Bing Liu. (2008). Opinion spam and analysis. In: Proceedings of the 1st ACMInternational Conference on Web Search and Data Mining.
6. Huayi Li, GeliFei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, Jidong Shao.(2017). Bimodal distribution and co-bursting in review spam detection. In: Proceedings of the26th International Conference on World Wide Web.
7. Arjun Mukherjee, Bing Liu,Natalie Glance. (2012). Spotting fake reviewer groups in consumerreviews. In: Proceedings of the 21st International Conference on World Wide Web (ACM).
8. MyleOtt, Yejin Choi, Claire Cardie, Jeffrey T Hancock. (2011). Finding deceptive opinionspam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of theAssociation for Computational Linguistics: Human Language Technologies.
9. VladSandulescu, Martin Ester. (2015). Detecting singleton review spammers using semanticsimilarity. In: Proceedings of the 24th international conference on World Wide Web(ACM).
10. Bondielli, A., &Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. 497, 38–55. https://doi.org/10.1016/j.ins.2019.05.035.
11. Cardoso, F., Cristina, A., & Garcia, B. (2019).Can Machines Learn to Detect Fake News? A Survey Focused on Social Media. 6, 2763–2770.
12. Han, W., & Mehta, V. (2019). Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation. Icii, 375–380. https://doi.org/10.1109/ICII.2019.00070.
13. Karimi, Hamid, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. "Multi-source multi-class fake news detection." In Proceedings of the 27th international conference on computational linguistics, pp. 1546-1557. 2018.
14. Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." arXiv preprint arXiv:1811.00770 (2018).
15. Zhang, Chaowei, Ashish Gupta, Christian Kauten, Amit V. Deokar, and Xiao Qin. "Detecting fake news for reducing misinformation risks using analytics approaches." European Journal of Operational Research 279, no. 3 (2019): 1036-1052.

16. Smitha, N., and R. Bharath. "Performance Comparison of Machine Learning Classifiers for Fake News Detection." In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 696-700. IEEE, 2020.
17. Ozbay, FeyzaAltunbey, and Bilal Alatas. "Fake news detection within online social media using supervised artificial intelligence algorithms." Physica A: Statistical Mechanics and its Applications 540 (2020): 123174.
18. Lu, Yi-Ju, and Cheng-Te Li. "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media." arXiv preprint arXiv:2004.11648 (2020).
19. Ibrishimova, Marina Danchovsky, and Kin Fun Li. "A machine learning approach to fake news detection using knowledge verification and natural language processing." In International Conference on Intelligent Networking and Collaborative Systems, pp. 223-234. Springer, Cham, 2019.
20. Kumar, S. (2022). Strategic management of carbon footprint using carbon collectible non-fungible tokens (NFTS) on blockchain. Academy of Strategic Management Journal, 21(S3), 1-10
21. Kumar, S. (2021). Review of geothermal energy as an alternate energy source for Bitcoin mining. Journal of Economics and Economic Education Research, 23(1), 1-12

22. Dr. Naveen Nandal, Dr. Aarushi Kataria, Dr. Meenakshi Dhingra. (2020). Measuring Innovation: Challenges and Best Practices. International Journal of Advanced Science and Technology, 29(5s), 1275 - 1285.
23. Nandal, N., & Nandal, N. (2019). BSCQUAL: A Measuring Instrument of Service Quality for the B-Schools . International Journal of Psychosocial Rehabilitation, Vol. 23, Issue 04, 1574-1589
24. Kaur, Sawinder, Parteek Kumar, and PonnurangamKumaraguru. "Automating fake news detection system using multi-level voting model." Soft Computing 24, no. 12 (2020): 9049-9069.
25. Kaliyar, Rohit Kumar, AnuragGoswami, and Pratik Narang. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach." Multimedia Tools and Applications 80, no. 8 (2021): 11765-11788.
26. Verma, Pawan Kumar, and Prateek Agrawal. "Study and Detection of Fake News: P 2 C 2-Based Machine Learning Approach." In Data Management, Analytics and Innovation, pp. 261-278. Springer, Singapore, 2021.
27. Nasir, Jamal Abdul, Osama Subhani Khan, and IraklisVarlamis. "Fake news detection: A hybrid CNN-RNN based deep learning approach." International Journal of Information Management Data Insights 1, no. 1 (2021): 100007.
28. Wang, William Yang. "" liar, liar pants on fire": A new benchmark dataset for fake news detection." arXiv preprint arXiv:1705.00648 (2017).