International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 08 | Aug - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

LLMv2: A Language Model with Holistic Hidden Markov Models and Machine Learning for Fake News Detection

[1]VENKATA HARISH BALAJI, VIT AP
[2]THEJESHWARI BALAJI, IIIT Raichur

Abstract – The usage of social media has expanded in recent years, allowing them to get news from around the world at any time. This in turn, is questioning the authenticity of the news that is being spread both globally and locally. Fake news such as misinformation, gossips is widely disseminated on social media having a negative impact on society and lives of the people. As a result, much study is being is carried out in order to detect them. The data can be clustered into smaller groups based on the type of news using a few learning approaches. A novel method has been proposed for prediction of the authenticity of the news of the LIAR dataset [1] using Logistic Regression and a boosting algorithm eXtreme Gradient Boosting (XGBoost) for efficacy, computational pace and performance of the model. This method detects fake news by analyzing the semantic and syntactic connections between sentences. Various graphs (like heat maps, bar charts) are plotted to show the distribution of the authenticity of news and also to compare the predicted result with the actual one. The proposed strategy addresses the effects of the hoax's global spread. People are hungry for information to defend themselves and others in a community where humans are confronting large-scale risks from harms. Some key traits such as Sentimental features, Content-based features, Frequency features, and Hybrid features (combinations of two or more features) are incorporated for early prediction of fake news spread via social media. The liar dataset is used to train the method and tested for accurate results. The experimental accuracy is found out to be 98%.

Key Words: fake news detection, social media, logistic regression, XG boost, natural language processing

1.INTRODUCTION

Every day, lot of stories are read on social media, some of which are legitimate, but the majority of them are not. Fake news consist of reports without verifiable facts, quotes, or sources as a result of this false or misleading information. Those stories are made up to manipulate or deceive readers. The term "fake news" refers to a way of thinking about how news is produced. Because false news stories spread faster than we can imagine on the internet, the number of fake news pieces has increased greatly in recent years via social media platforms such as WhatsApp, YouTube, Facebook, and others. The internet and technology have made it easier for fake news such as rumors and misinformation to circulate among individuals, resulting in a chaotic environment. In today's world, an innovative way for instantly detecting fake news is inevitable. Fake news may be quickly recognized with more than 90% accuracy using developing technologies such as Big data, Deep Learning, and Natural Language Processing (NLP)[2][3]. Many attempts to detect such news have been proposed, however they all rely on learning techniques. Here in this paper the trueness of the news is established on the LIAR dataset. This is publicly accessible dataset for detecting fake news from politifact.com [1]. Over the course of a decade, POLITIFACT.COM collected 12800 hand-labeled brief remarks in a variety of situations, including full analysis reports and links to source materials. This dataset can also be used to conduct fact-checking research. This new dataset is orders of magnitude larger than previous publicly available false news datasets of a similar nature. Many studies have been carried on the fake news detection [2-39] in order to verify the validity of the news. Bibek Upadhayay and Vahid Behzadan [4] applied deep learning architecture based on BERT-Base language model and acquired an accuracy of 70%. Akash Dnyandeo Waghmare, Girish Kumar Patnaik [5] used mNB in Blockchain for social media fake news detection and gained an accuracy of 95.20%. Harith H. Thannoon, Wissam H.Ali and Ivan A. Hashim [6] applied the classification algorithms like MLP, KNN, VG-RAM and SVM to detect the deception in facial expression. Chetana B. Thaokar and Jitendra Kumar Rout and Minakhi Rout [7] used sentimental features combined with metadata of the speaker which gives an accuracy of 75%. In response to those works, an enhanced ensemble strategy combining Logistic Regression, Extreme Gradient Boost (xGBoost) and Natural Language Processing (NLP) Techniques with tokens and word vectors [8] is developed to improve computation speed and accuracy. A. Abdulrahman et.al, [9] have used various machine learning and deep learning techniques to detect fake news and achieved an accuracy of greater than 81%. A. Roy et.al, worked with a deep ensemble framework for fake news detection and classification [10] showing an overall accuracy of 44.87%. This high level of precision is achievable with the proposed model (98%). This combination performs in semantic and syntactic analysis of the statements and obtain the accurate result for predicted and actual fake news..

2. Proposed Method

The workflow of the proposed method for fake news detection is shown in Fig.1. Data pre-processing techniques like punctuation removing, capitalization, lemmatization and removal of stop words are primarily carried out on the LIAR dataset s. The proposed framework builds on the extensions of machine learning methodologies. Tokenizers are used to create word cloud graphics as an extension of the models proposed by the research community (particularly Random Forest Classifiers (RFC), Support Vector Machine (SVM), and Decision trees). The method incorporates the semantic and syntactic connections between the statements then converting the sentences into vectors using vectorizers (Word2Vec, TF-IDF, and FastText). It also eliminates stop words, making

detection easier. This research employs ensemble methodologies as well as tuned linguistic feature sets such as Linguistic Inquiry and Word Count (LIWC) and Linguistic Text Analysis (LTA). The collected LIAR dataset is to be preprocessed for transforming raw data into a usable and efficient format. Various parts of the data are spotted to be missing or irrelevant and hence, data cleaning is performed in order to handle this component. It necessitate dealing with noisy data, missing data, and likewise. These characteristics include the data source, the format of the news presented, the number of stop words, the percentage of adjectives and pronouns used, the type of language used, and the accent, among others. The extreme gradient boosting, usually shortened as XG Boost is a scalable machine learning system for tree boosting that uses a tree-based ensemble machine learning algorithm [11-12]. It is widely used for both classification and regression, hence it is expected to serve the purpose well. It employs count vectorizer for extracting features from the statements and categorizing into percentage of authenticity of the news. Implementation of xGBoost enhances the performance for multiclass classification problems. The feature extracted dataset is shuffled for fair allocation and then split into training and testing dataset in the ratio 7:3. The ensemble model training is to be used to train the data and incorporates hyper-parameters tuning. Model evaluation is a crucial step in the creation of a model. Logistic Regression (LR) [13-17] model is to be assimilated to classify the fake news in binary output (either true or false). To establish a genuine classification of any unseen news, hypothesis of LR model manifests. The decision of the classification of Logistic Regression is based upon the function which is given as.

Every day, a vast number of stories are read on social media, some of which are legitimate, but the majority of them are not. Fake news consists of reports without verifiable facts, quotes, or sources, resulting in the spread of false or misleading information. These stories are often fabricated to manipulate or deceive readers. The term "fake news" has become a significant concern in how news is produced and consumed. As false news stories spread faster than ever on the internet, the number of fake news articles has increased significantly in recent years via social media platforms such as WhatsApp, YouTube, Facebook, and others. The internet and technology have made it easier for fake news, such as rumors and misinformation, to circulate among individuals, resulting in a chaotic environment.

In today's digital age, the necessity for an innovative and rapid method to detect fake news is inevitable. Emerging technologies such as Big Data, Deep Learning, and Natural Language Processing (NLP) can be used to quickly identify fake news with more than 90% accuracy . Several approaches have been proposed for fake news detection, many of which rely on machine learning techniques. This paper focuses on establishing the authenticity of news using the LIAR dataset, a publicly accessible dataset for detecting fake news sourced from politifact.com . POLITIFACT.COM collected 12,800 hand-labeled brief remarks over a decade in various contexts, including full analysis reports and links to source materials. This dataset is significantly larger than previous publicly available datasets of a similar nature and serves as a valuable resource for conducting fact-checking research.

Many studies have been conducted on fake news detection - to verify the validity of news articles. For instance, Bibek Upadhayay and Vahid Behzadan applied a deep learning architecture based on the BERT-Base language model, achieving an accuracy of 70%. Similarly, Akash Dnyandeo Waghmare and Girish Kumar Patnaik used mNB in Blockchain for social media fake news detection, gaining an accuracy of 95.20%. Another study by Harith H. Thannoon, Wissam H. Ali, and Ivan A. Hashim utilized classification algorithms such as MLP, KNN, VG-RAM, and SVM to detect deception in facial expressions. In contrast, Chetana B. Thaokar, Jitendra Kumar Rout, and Minakhi Rout employed sentimental features combined with the metadata of the speaker, which resulted in an accuracy of 75%.

In response to these studies, an enhanced ensemble strategy combining Logistic Regression, Extreme Gradient Boost (XGBoost), and Natural Language Processing (NLP) techniques with tokens and word vectors is developed in this paper to improve computation speed and accuracy. A. Abdulrahman et al. used various machine learning and deep learning techniques to detect fake news and achieved an accuracy greater than 81%. Another work by A. Roy et al. focused on a deep ensemble framework for fake news detection and classification, showing an overall accuracy of 44.87%. The proposed model in this paper, however, achieves a much higher level of precision with an experimental accuracy of 98%. This high level of precision is attainable through the combination of semantic and syntactic analysis of statements, resulting in accurate predictions of fake news.

3. CONCLUSIONS

The rapid spread of disinformation and false news on the social media can have adverse consequences. To combat this, the ensemble method using logistic regression and xGBoost would provide us with the clarity of the accuracy of the news. The focal point of the research lies in identifying and distinguishing among different categories of news such as barely true, true, false and so on. The proposed model aids in investigating and comprehending the effects of fake news on society by raising awareness of fake news propagators. The initial stage includes stopword removal, word vectorization, noun, pronoun count, average word length, meta-character count. In further stages, the articles' semantic and syntactic analysis is carried out and forecasted utilizing the aforementioned ensemble method. Hence, the accuracy of the proposed method is acquired to be 98%

REFERENCES

1. LIAR Fake News Dataset. Available at: [Kaggle](https://www.kaggle.com/datasets/csmalarkodi/liar-fake-news).

2. Apoorva Shete, Harshit Soni, Zen Sajnani, Aishwarya Shete, "Fake News Detection Using Natural Language Processing and Logistic Regression", Advances in Computing Communication Embedded and Secure Systems 2nd International Conference on, pp. 136-140, 2021.

3. Ahmad, Tahir, et al. "Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model." Applied Sciences, 12.3, 2022.

4. B. Upadhayay and V. Behzadan, "Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification," IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 1-6, 2020.

5. A. Dnyandeo Waghmare and G. Kumar Patnaik, "Social Media Fake News Detection using mNB in Blockchain," International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 1198-1204, 2022.

6. H. H. Thannoon, W. H. Ali, and I. A. Hashim, "Detection of Deception Using Facial Expressions Based on Different Classification Algorithms," 2018 Third Scientific Conference of Electrical Engineering (SCEE), pp. 51-56, 2018.

7. C. B. Thaokar, J. K. Rout, and M. Rout, "Classification of News Articles with Relational Multi Attributes using Machine Learning," 19th OITS International Conference on Information Technology (OCIT), pp. 216-220, 2021.

8. Lumbardha Hasimi, Aneta Poniszewska-Marańda, "Ensemble Learning-based Fake News and Disinformation Detection System", Services Computing (SCC) IEEE International Conference on, pp. 145-153, 2021.

9. A. Abdulrahman & M. Baykara, "Fake News Detection Using Machine Learning and Deep Learning Algorithms", International Conference on Advanced Science and Engineering (ICOASE), pp. 18-23, 2020.

10. A. Roy, K. Basak, A. Ekbal, & P. Bhattacharyya, "A Deep Ensemble Framework for Fake News Detection and Classification", ArXiv, abs/1811.0467, 2018.

11. V. Chandra Shekhar Rao, Pulyala Radhika, Niranjan Polala, Siripuri Kiran, "Logistic Regression versus XGBoost: Machine Learning for Counterfeit News Detection", Smart Technologies in Computing Electrical and Electronics (ICSTCEE) 2021 Second International Conference on, pp. 1-6, 2021.

12. Anggraina, R. Primartha, and A. Wijaya, "The Combination of Logistic Regression and Gradient Boost Tree for Email Spam Detection", Journal of Physics: Conference Series, Vol. 1196, 2016.

13. J. C. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, "Supervised Learning for Fake News Detection", IEEE Intelligent Systems, 2019.

14. S. Shabani and M. Sokhn, "Hybrid Machine-Crowd Approach for Fake News Detection", IEEE 4th International Conference on Collaboration and Internet Computing (CIC), IEEE, pp. 299–306, 2018.

15. S. Tyagi, A. Pai, J. Pegado, and A. Kamath, "A Proposed Model for Preventing the Spread of Misinformation on Online Social Media Using Machine Learning", Amity International Conference on Artificial Intelligence (AICAI), IEEE, pp. 678–683, 2019.

16. N. F. Baarir and A. Djeffal, "Fake News Detection Using Machine Learning", 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), 2021, pp. 125-130, 2020.

17. Wazib Ansar, Saptarsi Goswami, "Combating the Menace: A Survey on Characterization and Detection of Fake News from a Data Science Perspective", International Journal of Information Management Data Insights, Volume 1, 2021.

18. S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey", IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 436-441, 2018.

19. M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals", 22nd Conference of Open Innovations Association (FRUCT), pp. 272-279, 2018.

20. C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini and F. Menczer, "The Spread of Fake News by Social Bots", pp. 96-104, 2017.

21. R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica and A. N. Krishna, "Identification of Fake News Using Machine Learning", IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6, 2020.

This list provides a wide range of sources related to fake news detection, covering various methodologies and approaches.