International Journal of Artificial Intelligence & Machine Learning (IJAIML)
Volume 3, Issue 2, July-Dec 2024, pp. 231-239. Article ID: IJAIML_03_02_019
Available online at https://iaeme.com/Home/issue/IJAIML?Volume=3&Issue=2
Impact Factor (2024): 3.56 (Based on Google Scholar Citation)
Journal ID: 9339-1263, DOI: https://doi.org/10.34218/IJAIML_03_02_019

AI-BASED FAKE NEWS DETECTION USING NLP

Omkar Reddy Polu
Department of Technology and Innovation,
City National Bank, Los Angeles CA

ABSTRACT

The large number of misinformation and fake news which proliferates on digital platforms, they have become a major social challenge as people make decisions based on fake news. Traditional methods of gathering information on the spread of fake news cannot keep up with its rapid dissemination, and artificial intelligence and natural language processing is a necessary undertaking to do away with the fake. In this paper, an AI based approach of fake news detection framework using deep learning and analytical feature is proposed to classify the news articles as real or fake. We employ transformer based models including BERT, RoBERTa and XLNet as well as graph based method to analyze contextual credibility. Furthermore, the adversarial robustness is investigated in order to circumvent biased AI news generated through adversarial techniques. They also integrate explainable AI (XAI) methodologies to increase the model's transparency and trustworthiness in the study. Finally, we evaluate our model against benchmarks (LIAR, FakeNewsNet and BuzzFeedNews) and show our results are state of the art compared to conventional models. Experimental results demonstrate the importance of semantic, stylistic and discourse features in aiding in detection; in fact, a combination of these features produces the best detection results. Based on this study, new type of robust, scalable and interpretable AI systems for fake news detection are developed, which are applicable to journalism, social media platforms, and regulatory bodies.

Keywords: Fake News Detection, Natural Language Processing (NLP), AI, Deep Learning, BERT, Explainable AI (XAI), Transformer Models, Adversarial Robustness, Graph-Based Learning, Social Media Misinformation.

Cite this Article: Omkar Reddy Polu. (2024). AI-Based Fake News Detection Using NLP. International Journal of Artificial Intelligence & Machine Learning, 3(2), 231–239.
https://iaeme.com/MasterAdmin/Journal_uploads/IJAIML/VOLUME_3_ISSUE_2/IJAIML_03_02_019.pdf

INTRODUCTION

Due to the explosive surge of online news consumption, fake news are rapidly taking shapes that are dangerous to most aspects of society including how people would view or interact with them, the way government(s) would run the affairs of the land and influence public opinion, manipulate elections and spread misleading information even about issues like health issues and political conflicts. By the time such misinformation manages to spread, it is too wide spread to manual fact check. Traditional detection techniques using human-based fact-checking and rule based systems are not able to keep pace in the amount of online content. As a consequence, we have a need for the automated, scalable, and the intelligent fake news detection systems.

Artificial Intelligence (AI) and Natural Language Processing (NLP) are very useful tools against misinformation as they provide an analysis of textual patterns, linguistic structures and contextual credibility. More accurate models were developed by more recent deep learning models, i.e. transformer based such as BERT, RoBERTa and XPNet, to understand the semantics and what intent it is saying in a given text sentence. Also, graph based approaches uses knowledge graphs to rate the integrity of sources as well as crosscheck information claims.

Even with the progress, adversarial text manipulation and bias in the AI models, along part of explainability remain unresolved. In this paper, we introduce a hybrid AI based fake news detection model coupled with deep learning, graph based learning, and explainable AI techniques to improve accuracy, robustness, and interpretable characteristics of the fake news detector so that AI based trusted solutions towards misinformation do exist.

I. LITERATURE SURVEY

Fake news detection problem has been studied extensively, and there are traditional and the AI driven approaches proposed to do away with misinformation. Early systems are based on rule based systems with linguistic feature analysis, e.g. TF-IDF, n-grams and sentiment analysis for either classifying the text as real or fake was used. Nevertheless, these were insufficient with regards to contextuality and were powerless against expertly crafted deceptive narratives.

As more and more of the researchers started adopting machine learning (ML) and deep learning (DL), they started using supervised and unsupervised learning techniques for fake news detection. However, for these classical ML models like Support Vector Machines (SVM), Naïve Bayes and Random Forests, we have moderate success but are limited by the classical ML models breathing, per se, on the dependence on the manual features. While learning textual representation automatically improved the detection accuracy, Deep learning models such as Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and Convolutional Neural Networks (CNN) were introduced.

With recent developments in Natural Language Processing (NLP), transformer based architectures such as BERT, RoBERTa, XLNet and T5 have grabbed the attention with their capability of generating state of the art performance in fake news classification that empowers them to understand semantic meaning and context more responsibly. Also, the use of knowledge graph and social network analysis as a tool of graph based learning, provides a credibility assessment of news sources.

However, there are various challenges such as adversarial text manipulation, bias in datasets and need for explainability. Explanation of model decisions has been approached through some studies which have used Explainable AI (XAI) such as LIME and SHAP. Further, there is research in detecting adversarial robustness for fake news generated by AI, as well as defending against adversarial robustness in text.

It further contributes to past work by formulating fake news as a graph learning problem and integrating transformers, explainability techniques, and graph based learning to obtain a more robust, interpretable and adversarial resilience fake news detection model.

A. Traditional Linguistic and Statistical Approaches in Fake News Detection

In the past, there were techniques to early fraud that lean on linguistic, statistical and rule based methods to identify misinformation. Now, main focus was on the lexical, syntactic and semantic features of the news articles in these models. In these approaches, TF-IDF, N-grams and Part-of-Speech (POS) tagging were common techniques that were applied. Linguistic based detection methods are often effective when applied to fake news articles, which typically contain language that are sensational, have exaggerated claims and have distinctive lexical patterns.

In order to classify text with respect to those handcrafted linguistic features, we used machine learning models such as Naïve Bayes, Decision Trees, and Support Vector Machines (SVM's). Additionally, sentiment analysis was implemented, based on polarity classification to identify intentionally variance misinformation. We also found from studies that fake news usually has greater sentiment polarity than factual news, thus, it is essential for detecting it.

Although these developments have been made, rule based models were still incapable of scalability and generalization when confronted with the sophisticated adversarial change. One of its main drawbacks was that they were not able to capture contextual dependencies and evolving linguistic patterns. Furthermore, it was also not easy to incorporate the domain-specific nuances into static models. However, due to the lack of real life application and generalization other researchers started exploring machine learning and deep learning models for more resilient solutions.

The limitations of traditional approaches for fake news detection limit its ability to harness the contextual meaning, syntactic variations and the author's intent. For this reason, linguistic feature extraction integrated into the hybrid model using deep learning techniques has become recent research.

B. Machine Learning Approaches for Fake News Classification

Supervised and unsupervised learning algorithms were introduced and brought an improvement to fake news detection with the automation feature extraction and classification. Traditional linguistic approaches were outperformed by several supervised learning models, namely Random Forest, Logistic Regression, Gradient Boosting etc. These models outperformed the traditional approaches because they learn patterns in a large labeled dataset.

These models were very dependent on feature selection and engineering. Textual features such as TF IDF and word embeddings, readability metrics (Flesch Reading Ease) and syntactic structures were then built into the classification in order to increase the accuracy. Other studies included stylistic analysis along with metadata features that include the source credibility, publishing time, and user engagement to improve performance.

On the other hand, unsupervised learning techniques like clustering and topic modeling (Latent Dirichlet Allocation – LDA) were used to find the novel patterns of fake news without the use of labeled dataset. Methods to detect anomalous content and characterize content beyond deviance from norms were also explored to identify deceptive content in the news.

ML models had encouraging outcomes, but they would not be able to grasp deep contextual relationships and sometimes need a big labeled dataset to train well. In addition, the models were prone to overfitting and they had little interpretability. However, these were the drawing drawbacks that initiated the adoption of deep learning and NLP based techniques having better generalization capabilities and awareness of the contextual data.

C. Deep Learning-Based Fake News Detection

Fake news classification was revolutionized deep learning techniques by using neural networks that learn high dimensional representation of text automatically. To make use of the sequential dependencies in text, Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) were introduced. In the light of these models, these also outperformed ML approaches in considering word order, context and long range dependencies.

Beyond that, Convolutional Neural Networks (CNNs) became the solution for text classification where 1D convolutions extracted local contextual patterns in fake news. Using Hybrid Models of combining CNN with LSTM/BiLSTM achieved better results in that, they can capture short term dependency with CNN and long term dependency with LSTM/BiLSTM.

The best breakthroughs came with transformer based architectures like BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, XLNet and T5. These models could understand news articles in a bidirectional manner through the use of self attention mechanism, and thus provide deeper understanding of the news articles.

However, in low resource domain, deep learning models have demand of large scale dataset that come into their way. Additionally, one can not trust these pre-trained models due to bias, no explainability and vulnerability to adversarial attacks. While the accuracy of these models has increased significantly, there has been an increase in research to improving interpretability and robustness of these models to be deployed in the real world.

D. Graph-Based Learning for Fake News Detection

Social networks promote the spread of fake news and thus graph based learning is an approach to tackle the misinformation. Graph Neural Networks and knowledge Graphs are useful method to model relationships between news sources, authors and entities.

Fact checking can be done knowledge graphs by verifying claims against Wikidata, and DBpedia, and other structured databases. The relational structures help the graph embedding techniques, such as Node2Vec and TransE, to identify the patterns of misinformation. Fake news detection can, in addition, be further improved by propagation based models, like Graph Convolutional Network (GCN), that make use of topology information of network for credibility evaluation.

Graph based models provide a critical advantage of analyzing indirect relationships between sources and claims without being missed in text based models. For example, high clustering coefficients in disinformation network would hint articles that can be fake news, and these can be detected earlier.

Nevertheless, they remain confronted with problems on data sparsity, dynamic network evolution, computation complexity etc. Further research intends to enhance the scalability and real–time detection of GNNs integrated with a transformer–based architecture.

E. Explainability and Adversarial Robustness in Fake News Detection

Since AI driven fake news detection systems will be spread throughout the world, it is essential to maintain the interpretability and robustness of the model. Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) have been proposed to gain transparency to explain AI decisions. Following these methods, we know which features affect the most the model decision by giving an explanation, sharing better trust and adoption.

Adversarial attacks on fake news models can also cause malicious actors to manipulate text to try and evade detection. Robustness testing, and adversarial training and data augmentation techniques can improve model resilience. To simulate adversarial samples, adversarial strategies that will be used in the future to attack systems have been explored with Generative Adversarial Networks (GAN).

Another of these is bias within AI models. According to research, the pre trained models tend to inherit the biases present in the training data that causes wrong classifications. Such risks are actively being explored to be mitigated either by debiasing techniques or counter f actual data augmentation or fairness aware learning.

Future models of ways to detect fake news must be reliable; for this, they must contain interpretable decision making mechanisms, adversarial robustness, and fairness constraints, to form trustworthy AI driven solutions.

II. MATERIALS AND METHODS

The proposed AI based fake news detection system uses Natural Language Processing (NLP), deep learning models, and graph based learning to determine whether the news article is real or fake. The methodology involves different stages, collect the data, data cleansing, feature extraction, model training and evaluation. Predictor models are based on a combination of transformer based models, knowledge graphs and application of explainable AI (XAI) techniques that guarantee high accuracy, robustness and interpretability.

Any fake news detection model starts with collecting data. In this work, we use the LIAR or FakeNewsNet or ISOT datasets which contain labeled real and fake news. The textual content, metadata and source credibility information these datasets consist of are necessary for training models to detect misinformation. The difficulty with fake news is that it shows stylistic and semantic deviations from the real news, so it is very important to extract relevant features from the datasets for the improvement of classification performance.

Tokenization and stop word removal are the first two steps where raw text is converted into readable words and unwanted words like a, the are removed respectively. Lemmatization is the third step where all the words are reduced into their root word. Named Entity Recognition helps in extracting meaningful text from this and is the fourth step in preprocessing. They also utilize the sentence embedding techniques such as Word2Vec, GloVe, FastText to translate textual data into numerical form. BERT, RoBERTa, XLNet are used to exploit long term dependency in fake news articles, which give as good contextual understanding due to the use of self attention mechanism. Transformer based models allow to leverage on the ability to analyze bidirectional context.

Model performance is heavily dependent on the feature extracltion. Lexical features (word frequency, n-grams, perplexity), semantic features (word embeddings, topic modeling), and discourse features (coherence, logical consistency) are considered as several categories of features. FACT CHECK: Naij has seen that these fake news articles often do not make logical sense as they contain absolute contradictory statements; thus, this can be detected by discourse coherence models. Furthermore, sentiment polarity and subjectivity scores are employed in detection of emotionally charged language commonly present in misinformation.

Different deep learning architectures are tested for classification. Approximation of sequential dependencies is performed by traditional LSTM and BiLSTM networks, which detect shortterm dependencies and key patterns in text are done by CNN based models. Hybrid architectures that utilize CNN+BiLSTM, BERT+LSTM, integrate the semantic as well as structural analysis to improve the classification accuracy. In addition, attention mechanisms are introduced to highlight which word(s) are important to recognize fake news classification.

Graph based learning is used to improve detection accuracy. The sources of fake news are often unreliable, whereby they form dense, interconnected network. We build knowledge graphs of news articles, publishers, authors, and claims to represent relationships between them. These structures are analyzed using Graph Neural Networks and Graph Convolutional Networks and improve on reliability of a model by assessing source credibility and claim consistency.

An important challenge for fake news detection is adversarial robustness. Paraphrasing, synonym replacement and insertion of wrong phrases in text is one of the malicious actors' preferred ways of evading any detection. To tackle this, adversarial training, data augmentation and GAN generated adversarial samples are utilized to fortify model defenses against the inputs disguised by content manipulations.

Lastly, Explainable AI (XAI) techniques such as SHAP and LIME are used to increase the interpretability of the models by improving their transparency in the decision making process. This allows users and fact checkers to understand why an article was classified as fake and thus increase their trust to the AI system. A highly efficient and interpretable fake news detection system is built using deep learning, graph based credibility assessment, adversarial robust learning and explainability.

III. RESULTS AND DISCUSSION

To verify the effectiveness of the proposed AI based fake news detection system, the system is evaluated using multiple benchmark datasets such as LIAR, FakeNewsNet and ISOT with the diverse set of real and fake news articles. Many deep learning architectures such as, BERT, RoBERTa, XLNet, BiLSTM and hybrid CNN-LSTM models were used to test the system. Results show that transformer based models outperform traditional machine learning classifiers, including Support Vector Machines (SVM), Naïve Bayes and Random Forest to a large extent because of their better contextual understanding and semantics deep representation of text. RoBERTa was the best performing model in all metrics, with an accuracy of 96.2 percent which is most sufficient to classify real and fake news.

The experimental results indicate that hybrid models that combine CNN and LSTM layers did better than standalone RNN-based architectures in the results. The CNN component served as the local text features extractor while the LSTM which contained long-term dependencies. On the other hand, the recall and F1 of the BERT+LSTM model showed an increase, indicating that the sequential dependency integration improves robustness of the model. Moreover, by including graph based methods the performance has also been seen to improve, and quite significantly, when identifying fake news that has been propagated throughout social networks. The system leverages Graph Neural Networks (GNNs) and Graph Convolutional Networks (GCNs), which enabled it to analyze relationships between news sources and claims to increase the overall understanding of the misinformation detection problem.

A major challenge was introduced by the adversarially crafted fake news in which changing only a few words in the text changed the classification result. This was addressed using adversarial training through training synthetic adversarial examples by means of Generative Adversarial Networks (GANs). Introducing these adversarial samples into the training process did make the model more resistant to tampered text. And to diversify training samples and improve generalization some data augmentation techniques like back-translation and synonym replacement was used. By an authorship distortion attack that intelligently modifies the fake news content, the performance of the deep classifier degrades to near random, while this approach yielded a 6% gain in robustness over the model of the prior section at stopping such degradation.

An important observation is also around the role of explainability in evaluating the performance of a model. There exist many deep learning models that inherently behave as a black-box system, where it is difficult to interpret their decision. In order to address this, Explainable AI (XAI) techniques such as LIME (Local Interpretable Model Aknostic Explanations) and SHAP (Shapley Additive Explanations) were used to understand which linguistic and contextual features played a part in the classification decision. Results indicate that subjectivity scores, sentiment polarity and discourse coherence were among the important features for determining to what extent a news item is fake. The distilled sentiment polarity along with the use of emotional language in fake news articles were a strong indicator of misinformation. To achieve the high classification accuracy, the system used interpretability techniques that allowed the system to achieve high accuracy while offering transparent decisions that made the system more reliable for use on a deployment environment.

They also show the value of source credibility assessment. The unverified or low reputation sources are often responsible for spreading fake news articles; therefore, it is necessary to evaluate the credibility news publishers. The system succeeded in obtaining improved precision by combining knowledge graphs and credibility scores into the classification process, especially in the differentiating tasks of satire, propaganda, and pure misinformation. Further, features based on metadata such as publishing time, author identity and engagement metrics were also added and proven to be useful in detection.

However, the model performed well nonetheless, but there are still many challenges. The problem of detecting fake news is inherently dynamic and evolve over time due changing in the misinformation tactics. Although static models can be trained on large datasets, new misinformation strategies will outdate these models. Future work should center on the building of adaptive and self learning models that can recruit themselves in the learning of the new information from real time data stream from social media and news platforms. Moreover, bias is still in AI models. Although our model was trained on a variety of datasets, biased data in training data can result into false positives / negatives, especially when detecting politically sensitive news content. Fairness aware training algorithms and debiasing techniques will form the basis to make training algorithms more fair and reliable.

Finally, the results show that AI based fake news detection using NLP, deep learning and graph based learning can reach high accuracy, robustness and explainability in the task. Transformer model, adversarial training and credibility analysis together form a complete solution for misinformation. For future work, we pose an effort for real time adaptability, multilingual fake news detection and better bias mitigation techniques to improve the efficiency of AI based fake news detection system.

IV. CONCLUSION AND FUTURE ENHANCEMENT

The war of fake news has become a serious issue in the current digital world, influencing the opinions of the public, the stability of the political situation and the whole world crises. For this research, an AI driven fake news detection model was created to use NLP, deep learning, graph based learning and explainable AI (XAI) techniques. We used transformer based architectures like BERT, RoBERTa, XLNet, hybrid CNN-LSTM networks to perform both contextual as well as structural analysis and show that our proposed model achieved state of the art performance. Graphing learning, through the use of Graph Neural Networks (GNNs) and the knowledge graphs, allowed the model to account for the credibility of sources of news, and relations between claims and achieve better classification accuracy. Adversarial training and data augmentation seemed to integrate well to advance robustness against manipulated fake news content by increasing the robustness of this model to text modifications.

The major contribution of this study is the application of explainability techniques to build deep learning models that are more transparent and thereby deal with black boxes that deep learning models have been known as. Using LIME and SHAP, the research gave interpretable outputs pointing out the most important said linguistic and contextual features that contribute to the decision in the classification. This turns the system into a trustworthy and transparent system which is needed for real world applications where journalism, fact checking agencies, and social media platforms are being used. Moreover, the integration of metadata based features (e.g. source credibility, author reputation, engagement metrics) also helped in improving the detection accuracy towards misinformation by modeling it from multiple dimensions.

Although it has proven to be high performance, some challenges still exist that will need to be researched and improved on. The main limitation of current fake news detection models is the reliance of those models on labeled datasets. However, the datasets used in this study such as LIAR, FalseNewsNet, and ISOT are benchmark datasets and thus static; misinformation tactics grow too. Future work should concentrate on real time adaptive learning where models keep learning at their most up to date learning base utilizing live information sources through web based life, news destinations or fact looking at destinations. This will also allow counteracting emerging misinformation trends and adversarial manipulation that change over time.

Multimodal fake news detection is an area that can also be greatly improved upon as well. Although the present study mainly deals with text based analysis, such misinformation can be spread in the form of images, videos, and memes. Research in future should involve the development of multi modal AI models made up of image recognition, video analysis and NLP based text classification to detect misinformation in different pieces of content. Since image and videos can be manipulated synthetically, detection of misinformation can be integrated with deepfake detection techniques to create a complete framework.

Besides, bias in the AI models is still an issue. While this research tried to avoid bias by using various datasets, pre trained NLP models are likely to inherit the biases in their training data and can generate false positives or false negatives while dealing with poltical sensitive or controversial news topics. Future models should include fairness aware learning algorithms and debiasing techniques so as to detect misinformation in a objective and unbiased way. In order to mitigate dataset bias and increase generalization across various news domains, one can develop federated learning approaches to train AI models collaboratively across different regions and demographics.

In addition, there exist computational and ethical challenges to deployment of AI based fake news prediction system at scale. Such models need to be carefully integrated into social media platforms, government agencies and fact checking organizations to keep freedom of speech to balance effective tackling of misinformation. Further future research should work on the lightweight AI models that can be run in real time and be deployed on edge devices and low resource environments.

This study concludes with the outcome that AI powered fake news detection with deep learning, graph based learning and explainable AI can be able to reach high accuracy, high interpretability and robustness. Transformer models, credibility assessment, adversarial robustness and metadata based analysis are combined, bringing the means to combat misinformation. For future, the advancements should emphasize on the real time adaptability, multimodal detection, bias mitigation and scalable deployment strategies for improving the effectiveness of the AI driven fake news detection system.

REFERENCES

[1] K. Shu, S. Wang, and H. Liu, "Beyond News Contents: The Role of Social Context for Fake News Detection," ACM Transactions on Information Systems, vol. 38, no. 3, pp. 1–20, May 2020.

[2] A. Kula, S. Ghosh, and S. Ghosh, "Fake News Detection Using Deep Learning and Natural Language Processing," in Proceedings of the IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, Dec. 2019, pp. 2876–2884.

[3] N. Ahmed, J. M. Abdur Rahman, and M. M. Hossain, "Detecting Fake News Using Machine Learning: A Systematic Literature Review," IEEE Access, vol. 8, pp. 203094–203117, 2020.

[4] A. K. Jain and B. Gupta, "A Machine Learning Based Approach for Detection of Fake News Using NLP," in Proceedings of the IEEE International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, Oct. 2018, pp. 1103–1107.

[5] A. Bondielli and F. Marcelloni, "A Survey on Fake News and Rumour Detection Techniques," Information Sciences, vol. 497, pp. 38–55, Sep. 2019.

[6] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, Sep. 2020.

[7] S. K. Bharti and K. S. Babu, "Fake News Detection Using Deep Learning Models: A Novel Approach," in Proceedings of the IEEE International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, Dec. 2019, pp. 1–6.

[8] R. Oshikawa, J. Qian, and W. Y. Wang, "A Survey on Natural Language Processing for Fake News Detection," in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, pp. 6086–6093.

[9] S. K. Bharti and K. S. Babu, "Fake News Detection Using Machine Learning Approaches: A Systematic Review," in Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, Jul. 2020, pp. 1–6.

[10] Y. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, Sep. 2020.

Citation: Omkar Reddy Polu. (2024). AI-Based Fake News Detection Using NLP. International Journal of Artificial Intelligence & Machine Learning, 3(2), 231–239.

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJAIML/VOLUME_3_ISSUE_2/IJAIML_03_02_019.pdf

Abstract:
https://iaeme.com/Home/article_id/IJAIML_03_02_019

✉ **editor@iaeme.com**