International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

# DETECTING AI-GENERATED FAKE NEWS USING MACHINE LEARNING

**Mohammed Saifuddin[1], Mohammed Abbas Qureshi[2], Dr. Md Ateeq Ur Rehman[3]**
*[1]PG Scholar, Department of Computer Science and Engineering,*
*Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086*
*Email: Saof0910@gmail.com*
*[2]Assoc Professor, Department of Computer Science and Engineering,*
*Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086*
*Email: mohammedabbas09@yahoo.com*
*[3]Professor, Department of Computer Science and Engineering,*
*Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086*
*Email: mail_to_ateeq@yahoo.com*

**ABSTRACT**

The spread of AI-generated fake news poses increasing obstacles to the distribution of information as sophisticated large language models like GPT continue to evolve. Due to their limited ability to discern between real and fake news, traditional text classification techniques are unable to identify such content. This paper presents an MLP (Multi-Layer Perceptron) Classifier combined with Natural Language Processing (NLP) methods for identifying fake news produced by artificial intelligence (AI) in order to address this problem. Tokenisation, stop-word elimination, and vectorisation are preprocessing techniques used to extract significant features from textual data, which are subsequently fed into the MLP network. The classifier captures intricate linguistic characteristics that define fake news by utilising nonlinear activation functions and several hidden layers. To train and assess the system, a fresh dataset comprising 42 news categories was created using GPT-4. According on experimental data, the suggested MLP model outperforms conventional machine learning techniques by achieving strong F1 scores and dependable accuracy. These results demonstrate how MLP-based designs can improve the detection of fake news and protect the integrity of online information. [35], [4], [22],[23]

## 1. INTRODUCTION

The quick development of artificial intelligence in the current digital era has made it possible to produce extremely convincing AI-generated writing, such as news stories, social media posts, and other online content. Although these technologies, especially huge language models like GPT, have many advantages, they also present serious hazards since they make it easier for bogus news to proliferate. Fake news produced by AI has the potential to skew facts, sway public opinion, and erode confidence in online information sources. The complex linguistic patterns and contextual coherence generated by contemporary AI models frequently prevent traditional text classification and detection techniques from correctly identifying these fake articles. In order to overcome this difficulty, sophisticated methods that are able to identify minor syntactic and semantic indicators that distinguish real news from fake information are needed. The goal of this research is to create a reliable Multi-Layer Perceptron (MLP) classifier that combines Natural Language Processing (NLP) methods to efficiently identify bogus news produced by artificial intelligence. Tokenisation, stop-word elimination, and vectorisation are the preprocessing techniques used by the system to convert unprocessed textual data into machine-learning-ready numerical representations. With its many hidden layers and nonlinear activation functions, the MLP architecture is intended to capture sophisticated linguistic traits suggestive of fake news and model detailed patterns within textual data. A rich and demanding environment for training and evaluation is provided by a specific dataset created with GPT-4 that covers 42 different news categories. Using this dataset, the model gains a high degree of accuracy in differentiating between real and artificial intelligence-generated information. According to experimental data, the MLP-based strategy achieves good F1 scores across categories and consistent accuracy, outperforming conventional machine learning techniques. The study emphasises how crucial it is to combine NLP and deep learning methods to improve detection capabilities in the face of changing AI-generated false information. This study also emphasises how important automated methods are for maintaining the integrity of internet information and fostering public confidence. The suggested method offers a scalable and flexible solution for practical applications by taking into account both

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

linguistic complexity and contextual nuance. By focusing on proactive methods to stop the spread of false information, our work adds to the expanding field of AI-assisted fact-checking and digital content verification. Furthermore, as AI-generated material grows more complex, the results reinforce the necessity for ongoing development of sophisticated models. In order to ensure successful performance, this study offers a thorough method for identifying AI-generated fake news that combines thorough data pretreatment, strong machine learning, and domain-specific evaluation. [35], [4], [22],[23].

## 2. EXISTING SYSTEM OF PROJECT

- The Global-Local News Detection Model is a hybrid deep learning technique used in the current system for identifying AI-generated fake news. [4]
- This architecture combines TextCNN to extract local semantic patterns, BiLSTM to capture sequential dependencies, BERT for contextual word embeddings, and attention methods to highlight important textual elements. [21], [13]
- When combined, these elements allow the system to recognise minor distinctions between real and fake information by analysing both global and local linguistic clues in news stories. The model aims to maximise accuracy in identifying false information produced by AI by utilising a variety of feature extraction approaches.
- The way the system works is by preparing news items using common NLP techniques like tokenisation, embedding creation, and text cleaning. BERT is then used to extract deep contextual representations of words and phrases from the preprocessed data. While TextCNN layers extract fine-grained local information, BiLSTM layers capture long-term dependencies across sequences. [21] [13]
- By concentrating on the most instructive words or phrases, the attention mechanism improves this process even more.
- A classification layer that determines if a news piece is authentic or artificial intelligence-generated is then fed the combined results.
- Despite its effectiveness, the model's processing cost and complexity make it less appropriate for real-time or large-scale applications.

## 3. RELATED WORK

### 3.1 LANGUAGE MODEL

The development of NLP language models is reviewed in this subsection. Neural language models like Word2Vec and GloVe are discussed after more conventional methods like statistical learning techniques and n-gram models. The authors then discuss the evolution of Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs), emphasising their capacity to handle sequential input while pointing out their shortcomings in managing long-range dependencies. The section goes on to describe how the Transformer design transformed natural language processing (NLP) through self-attention processes, resulting in models like BERT and GPT. Improvements in contextual understanding, multilingual processing, event reasoning, and multimodal learning are highlighted in the review of more recent developments, such as ClarET, EventBERT, KnowModel, LaMini-LM, ChatGLM, GLM-130B, Flamingo, and EVA-CLIP. [22],[23],[21] [15] [13] [17] [18]

### 3.2 FACK NEWS DETEDTION

The research on detecting fake news is the subject of the next sections. It examines conventional statistics and machine learning techniques for detecting false information and talks about the shift to deep learning-based techniques. The authors examine how the ability of contemporary neural architectures to learn intricate semantic and contextual information from textual input has enhanced the efficiency of false news classification. They also draw attention to the increasing difficulty presented by AI-generated material and the requirement for more advanced detection frameworks that can differentiate between news produced by AI and that penned by humans. [4]

## 4. PROPOSED SYSTEM

- The suggested approach combines Natural Language Processing (NLP) methods with a Multi-Layer Perceptron (MLP) classifier to identify artificial intelligence (AI)-generated fake news. [35] [4]
- First, raw text data from news stories is pre-processed using techniques like TF-IDF or word embeddings, cleaning, tokenisation, stop-word removal, and vectorisation.
- Through this method, unstructured text is converted into structured numerical features that capture the news content's syntactic and semantic characteristics.
- The algorithm makes sure that the linguistic cues that differentiate real news from fake news are accurately captured by building a feature-rich dataset spanning several news categories. [4]

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- The MLP classifier, which has several hidden layers with nonlinear activation functions, receives the features once they have been retrieved. [35]
- The model uses gradient descent and backpropagation to optimise weights in order to understand intricate correlations between textual elements and news authenticity.
- Deep contextual patterns that are frequently overlooked by conventional classifiers can be captured by the system thanks to the ensemble of hidden neurones.
- The system can accurately identify fresh inputs as either real or artificial intelligence-generated fake news after training on a varied dataset created with GPT-4. [4] [22],[23]
- The suggested approach is a solid contender for practical fake news detection systems since it guarantees scalability, robustness, and enhanced interpretability. [4]

## 5. MODULES

### 1. Data Collection:

In order to create a wide and well-balanced dataset, this module gathers information from both real and artificial intelligence-generated news sources. While fake news samples are produced using GPT-4 in 42 distinct categories, real news articles are collected from reputable web sources. To improve model generalisation, the gathering process guarantees a broad range of subjects, including politics, sports, technology, and entertainment. Every record has source information, category designations, and textual content. A solid basis for training and assessing the false news detection model is provided by the gathered data. Duplicates and unnecessary information are eliminated by proper data cleaning and filtering. [4] [22],[23]

### 2. Dataset:

The dataset simulates real-world instances of disinformation by combining genuine and artificial intelligence-generated news stories. It is classified into two primary groups: genuine and fake. The article title, content, and category tag are all included in each entry, offering contextual and semantic characteristics for analysis. To guarantee diversity and representativeness, the dataset spans 42 news domains. It is the primary input used to train, test, and validate the MLP classifier. To prevent bias in model learning and improve prediction accuracy, the dataset is meticulously balanced and arranged. [35]

### 3. Data Preparation:

In order to transform unprocessed text into useful numerical characteristics, this subject focuses on preparing textual data using Natural Language Processing (NLP) approaches. Tokenisation, stop-word elimination, and vectorization—which uses TF-IDF or Word2Vec to represent words numerically—are among the steps. Punctuation, special symbols, and superfluous whitespace are eliminated in order to normalise the cleaned text. To ensure accurate evaluation, the data is then separated into training and testing sets. The linguistic and semantic clues that separate authentic news from content produced by artificial intelligence are captured through feature extraction. This phase guarantees that the input data is organised, clean, and prepared for model training. [17]

### 4. Model Selection:

The Multi-Layer Perceptron (MLP) classifier is chosen as the main model in this module because of its capacity to identify intricate and nonlinear relationships in data. The MLP learns linguistic and semantic patterns suggestive of AI-generated fake news using a number of hidden layers and nonlinear activation functions. Hyperparameter tweaking, which includes changes to learning rate, batch size, and hidden layer dimensions, optimises the model's architecture. Compared to conventional machine learning algorithms, the selection of MLP guarantees resilience and flexibility. The basis for precise and scalable fake news classification is established by this module. [35] [4]

### 5. Analyze and Prediction:

In this module, preprocessed data is used to train the MLP model, and its prediction performance is evaluated. In order to comprehend the linguistic distinctions between real and AI-generated samples, the model learns from both during training. After being trained, the system uses linguistic characteristics to determine if a new article is authentic or fraudulent. Key words, sentence structures, and tone patterns that affect the prediction are identified as part of the investigation. To analyse model findings and comprehend predicted trends, visualisation techniques are utilised. This module guarantees that the model can use unseen data to produce precise, data-driven decisions. [35]

### 6. Accuracy on Test Set:

This module uses common measures like accuracy, precision, recall, and F1-score to assess the trained model's performance on the test dataset. The superiority of the MLP classifier is validated by comparing its findings with those of conventional models such as

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Naïve Bayes and Logistic Regression. The classifier's dependability and capacity to differentiate between authentic and fraudulent news are evaluated using confusion matrices and ROC curves. The assessment aids in figuring out how well the model adapts to new textual data. Prior to deployment, this module makes that the system is reliable and consistent. [35]

**7. Saving the Trained Model:**

The trained MLP model is saved for later use using serialisation libraries like pickle or joblib after it has reached a suitable level of accuracy. This makes it possible for the model to load quickly and detect bogus news in real time without retraining. Automated fact-checking tools, browser extensions, and web apps can all include the saved model. To guarantee reproducibility, version control and appropriate documentation are upheld. This module makes sure the detection system is scalable, effective, and ready for deployment in practical applications. [35]

## 6. RESULT AND DISCUSSION

The proposed system for detecting AI-generated fake news was evaluated using a dataset consisting of **4,200 news articles** generated across **42 categories**. The dataset included both genuine news articles and AI-generated fake news created using GPT-4. The data was divided into **80% training data** and **20% testing data** for performance evaluation.

The model employed a **Multi-Layer Perceptron (MLP) classifier** combined with Natural Language Processing techniques such as tokenization, stop-word removal, and TF-IDF vectorization for feature extraction.


Fig.1 Login Interface


Fig.2 Illustrates Input interface where user can enter The news for Verification.


Fig.3 Detection showing output of real news.

## 7. CONCLUSION

The suggested study effectively illustrates how an MLP-based classifier combined with NLP methods can identify bogus news produced by artificial intelligence. The approach precisely captures intricate linguistic patterns and semantic clues by preprocessing textual data using tokenisation, stop-word removal, and vectorisation. The findings show that when it comes to detecting fake news, the MLP classifier outperforms conventional machine learning techniques. The recently created dataset, which included both AI-generated and human-written pieces, offered a solid basis for assessment and training. The system's robustness and dependability were confirmed by its excellent accuracy and outstanding F1-scores. Overall, in the age of sophisticated AI content creation, this study demonstrates the promise of deep learning in maintaining information integrity, fostering digital trust, and battling false information. [35] [4]

## 8. FUTURE ENHANCEMENT

By using cutting-edge deep learning architectures like Transformers and BERT for better text comprehension, the suggested false news detection system can be improved in the future. Detection accuracy can be improved by using multimodal analysis that incorporates photos, videos, and social media

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

metadata. Additionally, the system may be extended to handle multilingual datasets, making it possible to identify bogus news in many languages and geographical areas. AI-generated false information can be quickly identified as it propagates online by using real-time news stream analysis. Explainable AI (XAI) methods can also be used to improve the transparency and interpretability of the prediction process. The solution can be made available to the public and media organisations through cloud-based deployment and API integration. As language models develop, ongoing retraining with freshly created AI content will help sustain high accuracy. Together, these improvements would increase the system's efficiency, scalability, and adaptability for upcoming digital ecosystems. [21] [15]

**REFERENCES**

R. Zhao and T. Shi, "The impact of large language models on public security intelligence work and countermeasures research," J. Big Data Comput., vol. 2, no. 1, pp. 91—103, Mar. 2024.

[2] N. Bontridder and Y. Poullet, "The role of artificial intelligence in disinformation," Data Policy, vol. 3, p. e32, Jan. 2021.

[3] S. Kreps, R. M. McCain, and M. Brundage, "All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation," J. Exp. Political sci., vol. 9, no. 1, pp. 104-117, 2022.

[4] M. R. Kondamudi, S. R. Sahoo, L. Chouhan, and N. Yadav, "A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches," J. King Saud Univ.-Comput. Inf.' Sci., vol. 35, no. 6, Jun. 2023, Art. no. 101571.

[5] A. Orhan, "Fake news detection on social media: The predictive role of university students' critical thinking dispositions and new media literacy," Smart Learn. Environments, vol. 10, no. l, p. 29, Apr. 2023.

[6] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[7] L. Hu, S. Wei, Z. Zhao, and B. wu, "Deep learning forfakenews detection: A comprehensive survey," Al Open, vol. 3, pp. 133—155, Jan. 2022.

[8] H. Zhao, H. Chen, T. A. Ruggies, Y. Feng, D. Singh, and H.-J. Yoon, ,'Improving text classification with large language model-based data augmentation," Electronics, vol. 13, no. 13, p. 2535, Jun. 2024.

[91 K. Shu, A. Bhattacharjee, F. Alatawi, T. H. Nazer, K. Ding, M. Karami, and H. Liu, "Combating disinformation in a social media age," WIREs Data Mining Knowl. Discovery, vol. 10, no. 6, p. 1385, Nov. 2020.

[10] E. Shushkevich, M. Alexandrov, and J. Cardiff, "Improving multiclass classification of fake news using BERT-based models and ChatGPTaugmented data," Inventions, vol. 8, no. 5, p. 112, Sep. 2023.

[11] A. Naitali, M. Ridouani, F. Salahdine, and N. Kaabouch, "Deepfake attacks: Generation, detection, datasets, challenges, and research directions," Computers, vol. 12, no. 10, p. 216, Oct. 2023.

[12] P. Dhiman, A. Kaur, D. Gupta, S. Juneja, A. Nauman, and G. Muhammad, "GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection," Heliyon, vol. 10, no. 16, Aug. 2024, Art. no. e35865.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[141 A. Rakhlin, "Convolutional neural networks for sentence classification," GitHub, vol. 6, p. 25, Jan. 2014.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, Jun. 2017, pp. 5998—6008.

[16] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J., vol. 27, no. 3, pp. 379-423, Jul. 1948.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proc. Adv. Neural Inf Process. syst., vol. 26, Dec. 2013, pp. 3111-3119.

[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proc. Conf Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1532-1543.

[19] J. Elman, "Finding structure in time," Cognit. Sci., vol. 14, no. 2, pp. 179-211, Jun. 1990.

[20] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Netw., vol. 5, no. 2, pp. 157-166, Mar. 1994.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, vol. 1, Minneapolis, MN, USA, Jan. 2018, pp. 4171-4186.

[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., 2018.

[23] T. B. Brown et al., "Language models are few-shot learners," in Proc. Adv. Neural In/ Process. syst., Jan. 2020, pp. 1877-1901.

[24] Y. Zhou, T. Shen, X. Geng, G. Long, and D. Jiang, "ClarET: Pretraining a correlation-aware context-to-event transformer for event-centric generation and classification," in Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, 2022, pp. 2559-2575.

[25] Y. Zhou, X. Geng, T. Shen, G. Long, and D. Jiang, "EventBERT: A pretrained model for event correlation reasoning," in Proc. ACM Web Conf. , Apr. 2022, pp. 850-859.

[26] Y. Zhou, X. Geng, T. Shen, J. Pei, W. Zhang, and D. Jiang, "Modeling event-pair relations in external knowledge graphs for script reasoning," in Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP, Jan. 2021, pp. 4586-4596.

[27] S. Dudhmande, S. Golliwar, A. Bhagwat, R. Ghiya, and A. Bhade, "Textual compression using lamini-LM," Int. Res. J. Adv. Eng. Manage. (IRJAEM), vol. 2, no. 5, pp. 1536-1540, May 2024.

[28] H. Lei, X. Liu, G. Niu, Y. Zhou, and Y. Zhou, "Generative Al authorship verification based on ChatGLM," in Proc. Work. Notes CLEF, 2024, PP. 1-6.

[29] A. Zeng et al., "ChatGLM: A family of large language models from GLM130B to GLM-4 all tools," 2024, arXiv.•2406.12793.

[30] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in Proc. Conj: Neural Inf. Process. Syst., Jan. 2022, pp. 1—21.

[311 Z.-J. Yang, T.-F. Feng, and X.-G. Wu, "On the possibility of mixed axion/neutralino dark matter in specific SUSY DFSZ axion models," 2023, arXiv.•2303.11645.

[32] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in Proc. 20th Int. Conf. World Wide Web, Mar. 2011, pp. 675—684.

[331 R. Labaca-Castro, "Generative adversarial nets," in Proc. Adv. Neural Inf.' Process. syst., vol. 27, Jan. 2023, pp. 73-76.

[34] X. Li and Y. Zhou, "Disentangled and robust representation learning for bragging classification in social media," in Proc. IEEE Int. Conj: Acoust., Speech Signal Process. (ICASSP), Jun. 2023, pp. 1-5.

[35] K. M. DSouza and A. M. French, "Fake news detection using machine learning: An adversarial collaboration approach," Internet Res., vol. 34, no. 5, pp. 1664—1678, Sep. 2024.

[36] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," Int. J. Inf Manage. Data Insights, vol. l, no. l, Apr. 2021, Art. no. 100007.

[37] J. Alghamdi, Y. Lin, and S. Luo, "The power of context: A novel hybrid context-aware fake news detection approach," Information, vol. 15, no. 3, p. 122, Feb. 2024.

[38] R. Shao, T. Wu, and Z. Liu, "Detecting and grounding multi-modal media manipulation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 6904-6913.

[39] F. Shan, H. Sun, and M. Wang, "Multimodal social media fake news detection based on similarity inference and adversarial networks," Comput., Mater: Continua, vol. 79, no. 1, pp. 581—605, 2024.

[40] S. Wachter, B. Mittelstadt, and C. Russell, "Do large language models have a legal duty to tell the truth?" SSRN Electron. J., vol. 11, no. 8, 2024, Art. no. 240197.

[41] R. Misra. (2022). News Category Dataset. [Online]. Available: https:// huggingface.co/datasets/heegyu/news-category-dataset