Bharathi Panduri, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

AI-Powered Misinformation Detection System Attrition

Bharathi Panduri[1], P K Abhilash[2], Dr. Y J Nagendra Kumar[3], Kaliveni Naveen[4], Alluru Manoj[5] , Patha Shiva Anurag[6]

Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology (GRIET), affiliated with JNTUH, Bachupally, Hyderabad, India

Abstract- The widespread dissemination of misinformation through social media, news platforms, and messaging applications has become a major challenge in the digital age, particularly in India. False and misleading information can trigger public panic, financial fraud, health risks, and social instability. Existing fact-checking solutions are often manual, time-consuming, or inaccessible to common users, creating a need for an intelligent and real-time misinformation detection system. This project, Verify & Learn, presents an end-to-end solution that leverages Artificial Intelligence (AI), Machine Learning (ML), and Generative AI (GenAI) to detect, explain, and educate users about misinformation. The system analyzes digital content such as news headlines, articles, social media posts, URLs, and messaging text using machine learning and natural language processing models. These models classify content into categories such as true, misleading, or false across multiple domains including health, politics, finance, science, social media, and online scams. To improve reliability and transparency, the system integrates trusted external sources through fact-checking and news APIs, enabling evidence-based verification with proper citations. Generative AI plays a crucial role in enhancing explainability and user understanding by converting analytical results into clear, human-readable explanations. It also generates educational insights, domain-specific awareness tips, and short learning content that help users recognize common misinformation patterns and manipulation techniques. The proposed system is implemented as a web application and a browser extension, allowing real-time verification during browsing and deeper analysis through detailed reports and shareable PDFs. By combining predictive ML models with explainable and educational Generative AI, this project promotes digital literacy, informed decision-making, and responsible information sharing. The solution demonstrates a scalable and user-centric approach to combating misinformation in real-world digital environments.

Keywords: Artificial Intelligence, Machine Learning, Generative AI, Misinformation Detection, Fake News Analysis, Natural Language Processing, Fact-Checking Systems, Explainable AI, Digital Literacy, Browser Extension, Online Scams Detection.

I. INTRODUCTION

The increase in digital communication methods (social media, online news portals, and messaging applications) have drastically changed the manner that information is created and consumed. This also has contributed to the significant increase in the amount of misinformation that is circulated, resulting in public panic, health risks, financial fraud, and societal instability. Soroush Vosoughi's research shows that false information is spread at faster rates and to wider audiences than true information due to the novelty and emotionality of false information. As such, Kai Shu has documented how the use of machine learning and data mining technologies are becoming increasingly important for detecting misinformation. The development of deep learning methods and transformer-based architectures has improved the effectiveness of systems for processing and classifying textual data, making these techniques vital for combating misinformation challenges.

Systems for identifying erroneous information (in this case, "misinformation") have improved quite a bit since their inception; however, still they are limited in their usefulness as described below. Many current solutions use a binary type of output (i.e., are categories classified as 'True' or 'False') and typically do not provide users with a reasonable way to determine how their input correlates with the output classification or decision made. Kai Shu argues the absence of interpretability has decreased trust in these types of systems, thereby limiting their use and overall success in academia and institutional uses. As noted by Soroush Vosoughi, explanations of the rapid spread of misinformation is primarily attributed to various human behavioral traits; however, most detection models fail to consider the role of these behaviors when classifying data with respect to an example of whether it is true or false. For example the Transformer models built and tested by Jacob Devlin perform well when classifying text, but the output decisions do not inherently provide sufficient detail for the user to develop reasonable inferences about how to interpret an AI-generated prediction vs. user observations.

This project aims to create an AI-based misinformation detection system that classifies information (true, false, or misleading) and gives users detailed, easy-to-understand explanations for their results. It is built on top of advanced NLP and transformer models based on Jacob Devlin's work to achieve high accuracy. The system uses Generative AI to create documents containing human-readable explanations, tips for awareness, and educational content to solve the issues of interpretability discussed by Kai Shu, and by combining the detection with an explanation element, the system will help to reduce the dissemination of misinformation, which is a concern raised in Soroush Vosoughi's research, and to encourage responsible sharing of information and digital literacy among users.

The goal of this study is to develop and implement a method for detecting false information in text-based content. This includes news articles, social media posts, and messages across a variety of domains (health, political, and financial). The proposed system consists of a web application as well as a means for users to access the information it provides via their web browsers, allowing them to see the results in real-time and interact with them while doing so. Although the methodology uses transformer-based models as proposed by Jacob Devlin and addresses interpretability issues identified by Kai Shu, it does not provide a means for identifying multimedia sources (e.g., images or videos). Furthermore, while the system enhances user comprehension through the use of explanations, it does not provide an adequate model of the intricate behavioral dynamics exhibited by humanity in terms of spreading false information (as discussed by Soroush Vosoughi). These limitations specify the context and limits of this proposed research project.

II. LITERATURE SURVEY

1. In 2024, Ahmad et al. did a study on finding hate speech in Arabic. They made a new multi-class dataset with 403,688 labelled tweets. By utilizing different text representation methods

such Word2Vec, TF-IDF, and AraBERT, the authors 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS) | 979-8-3315-29765/24/$31.00©2024IEEEDOI:10.1109/COMPAS60761.2024.10795890 Authorized licensed use limited to: Zhejiang University. Downloaded on February 20,2025 at 13:16:22 UTC from IEEE Xplore. Restrictions apply. tested several techniques of ML such as SVM, LR, Naive Bayes, Random Forest, AdaBoost, XGBoost and Catboost. CatBoost reached best accuracy of 57% with the help of Word2Vec. The main restriction was the possible absence of coverage of all hate speech variants among several Arabic dialects, therefore compromising the generalizability of the models.

2. Rohera et al. evaluated many machine learning models in a thorough investigation on false news categorization methods. Drawn from many sources including interviews, news releases, and speeches, they assembled a dataset of 7796 articles tagged with truthfulness, titles, and context. There were 80% for training and 20% for testing out of the dataset. The study examined many models: LSTM, Random Forest, SVM (with a passive-aggressive method), and Naive Bayes. With 92.42%, the LSTM model among these attained the best accuracy. This research has a clear strength in its careful comparison of several models; yet, the study might be strengthened even further by enlarging the dataset and including more varied sources to raise the generalizability of the outcomes.

3. Amer et al. did a detailed study on false news identification using several machine learning and deep learning models. To guarantee an all-around evaluation of model performance, they trained and assessed the classifiers using five-fold crossvaluation. For word embeddings, the dataset was glove.840B.300d; they also investigated contextual and statistical aspects. The machine learning models that were explored comprised Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Naive Bayes (NB). At the same time deep learning models have been also explored that include Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from transformers (BERT). Reaching 99.7%, the BERT model among these attained the best accuracy. With improvements ranging from 2-15% in accuracy, the study showed that contextual elements greatly enhanced model performance relative to statistical features. The study's shortcomings included the need of ongoing retraining to keep up with changing fake news generating tools and the absence of news source verification notwithstanding the strong results. Dealing with these constraints in next studies will help to improve the strength and relevance of the suggested models even more.

4. Piya et al. did a study using both English and Bengali datasets to create a bilingual model for finding fake news online. For English content, they utilised the ISOT Fake News dataset; for Bengali material with 50,000 annotated entries, they Six distinct ML techniques. With bigrams, the Linear SVC among these models obtained the best accuracy of 93.29%. The imbalanced dataset and the lack of annotated data in Bengali caused restrictions on the research even with the remarkable outcomes. Positively, this emphasises the possibility for major advancements by gathering additional data and experimenting with new deep learning models, therefore improving the real-time efficiency and general performance of the mode.

5. In 2021, S. Verma and N. Jain presented a hybrid method in Attrition Prediction Using Hybrid Machine Learning Models that combined the SVM and Random Forest classifiers to reach an accuracy of over 92%. The hybrid model maximized the performance of both classifiers and used SVM's boundary placement and Random Forest's capabilities of dealing with features. In another paper Feature Engineering Techniques for Employee Churn Prediction, K. Tanwar and S. Agarwal (2021) stressed the importance of inclusion of new

features called Tenure Balance and Income-Satisfaction Index, which raised the prediction accuracy by around 6%. In another study, S. Thakur and P. Nair (2021) conducted a comparative evaluation of the classifiers Logistic Regression, Decision Tree, and Random Forest, with Random Forest achieving around 95% accuracy when predicting employee attrition.

6. Khan et al. used the FakeNewsNet dataset to carry out a comparative investigation analyzing the applicability of different machine learning models for fake news detection. As deep learning models, they not only analyzed convolutional neural networks (CNN), long short-term memory (LSTM), Bidirectional LSTM (Bi-LSTM), Convolutional LSTM (CLSTM), Hierarchical Attention Network (HAN) neural network, and Convolutional Hierarchical Attention Network (Conv-HAN), but also a whole range of traditional classifiers such as SVM, logistic regression, decision trees, Naive Bayes, or k-NN. They further appraised new generation pre-trained language models such as ELMo, BERT, RoBERTa, DistilBERT, or ELECTRA. With a 96%, RoBERTa came out as having the best accuracy. The study underlined the durability and great performance of pre-trained models, particularly in situations with little training data, therefore proving their efficiency. Notwithstanding the computational complexity and possible overfitting problems with some models, the work offers a thorough investigation that might direct next research and pragmatic uses in fake news identification.

7. Shaikh et al. used several classification methods in order to identify bogus news. Their dataset consisted of both real and fake news items, drawn from a mix of online and handcrafted initiatives. Models including SVM, Naïve Bayes, and Passive Aggressive Classifier were tested. Out of all these, the SVM model produced the best accuracy of 95.05%. The great training time needed for the SVM model compared to the other classifiers, which could be less effective for bigger datasets, was a clear restriction of the research, though.

8. Sharma et al. shown the research about bogus news utilizing machine learning techniques. Their method combined TF-IDF Vectorization with Natural Language Processing (NLP) methods like CountVectorization. Using a Passive Aggressive Classifier, they classified news items using these methods. Although the particular dataset name was not stated, the materials for their study came from online news items. After training their model using four distinct classifiers, they chose the best-performing one for eventual use. In regard to bogus news detection, the Passive Aggressive Classifier attained 92.74% accuracy. One drawback of their research, nevertheless, was that it mostly concentrated on the technical execution without thorough examination of the bias in the dataset or the practical relevance of their mode.

9. J. and Meenakowshalya A. investigated the approaches for recognizing false information via n-gram analysis and various ML techniques. Considering six models, which are Support Vector Machines (SVM), Linear Support Vector Machine (LSVM), K Nearest Neighbour (KNN), Stochastic Gradient Descent (SGD), Decision Tree (DT) and Logistic Regression (LR). They used a dataset from Kaggle and obtained an accuracy of 93.5%, the SGD model produced the best accuracy the research revealed. Random Search CV performance tuning raised the accuracy to 94.2%. The restricted availability of resources and the declining accuracy with increasing n-gram size defined the limits of this research.

III. METHODOLOGY & SYSTEM ARCHITECTURE

System Overview

The new AI system that can tell when people are giving out false or misleading information uses advanced technology to find, understand and explain the source of the information instantly. The

system will use state of the art techniques such as NLP, generative AI, and transformer models in order to provide the user with an accurate classification of the content they provided, and a user-friendly explanation of how their content was determined to be accurate or not accurate. This system will be developed as both a web app and as a browser extension and will allow anyone to check if any text, such as headlines, social media posts or text messages, are accurate or not instantly.

Data Preprocessing

The system will allow users to provide text input from a variety of sources, including URLs, user provided content, and social media posts. Once a user has submitted their input, the text will be pre-processed using various techniques, including tokenizing the text, removing stop words, standardizing the format of the text, and encoding the text so that it can be analyzed by the classification model. Following the pre-processing of a user's input, the processed text will be provided to the classification model for analysis.

Misinformation Detection Model

The base for this system is a BERT/Distil-BERT transformer-based NLP model to classify content in terms of accuracy. For example, determine whether something is true, false or misleading. Because of their ability to comprehend contextual relationships between words/phrases in a body of text, these models provide much more accurate predictions than traditional machine-learning models. Training on labeled datasets with various forms of misinformation from multiple domains, including but not limited to: health, politics, and finance, ensures that the model will have strong domain-adaptive performance characteristics.

Explainability using Generative AI

Another component of the system that addresses some of the challenges associated with traditional supervised machine-learning black box models is a generative AI module used to explain each classification decision. This module creates human-readable text that describes what specific wording or images caused the content to be classified as false or misleading and also identifies the technique used to produce the misinformation (for example, common techniques used to create misinformation include exaggeration, emotional manipulation, or unsupported claims). By providing human-readable text that explains the classification for each piece of content, this module will also increase transparency and build trust between users and AI.

Integration with External Knowledge Sources

In addition to providing transparency, the model integrates external fact-checking sources (such as fact-checking APIs) and trusted news sources to increase reliability and provide more accurate classifications. Integrating external sources of evidence allows the model not only to rely on internal predictions but also to validate predictions from data found in the real world. The result is an increase in credibility and reducing false positives.

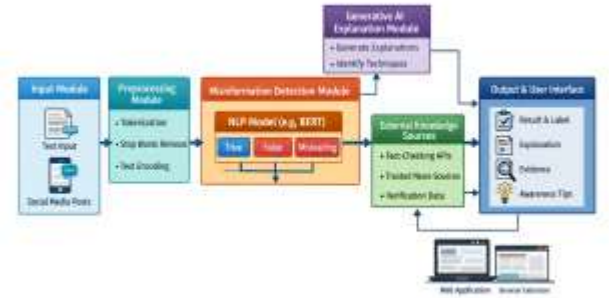
User Interface and Deployment

This project will be implemented as a web application to provide interactivity and ease of use by users and will be based on streamlit framework which is a part of the python programming language; a web browser (extension) will also be created so users can have their content analysed simultaneously as they browse. The system provides results to users using a clear and understandable format such as classification labels, reasoning for classification, and educational quotes about misinformation to provide the same access to all users whether they have a background in programming or not.

Expected Workflow

The process starts when a user inputs their text into the system after which it goes through a process of cleaning and obtaining features from that text before being fed into the transformer based model to produce a prediction for the user. This prediction will then go to the generative ai section of the systems output module which will produce both the reason(s) for making the prediction and/or any awareness tips for the user. The system provides users with all outputs from various external resources and provides users with all results in a manner that is complete and informative.

Architecture



The proposed AI-powered misinformation detection system architecture diagram shows how data is processed from input through the various components of the system until it is output in a useful way. Each of the four major blocks of the architecture diagram (user input, pre-processing module, misinformation detection module (core of the system), and output) performs a specific function that enables the system to accurately detect misinformation while providing a meaningful explanation of why something is misinformative.

Misinformation detection system architecture begins with the Input Module, where the user provides input (input data) to the system in the form of text (e.g., news article, social media post, or URL). The input is then passed on to the Preprocessing Module, where the data is cleaned and formatted for analysis purposes. The Preprocessing module provides operations for tokenizing, removing stop words, normalizing, and encoding text data, thus allowing raw text data to be structured into a format compatible with machine learning algorithms.

After passing through the Preprocessing Module, data is sent to the Misinformation Detection Module, which is the most important part of the system. The Misinformation Detection Module consists of an NLP transformer model (e.g., BERT or DistilBERT) that analyzes the context of the text and classifies it into one of the following three categories: true, false, and misleading. Using deep learning techniques, the NLP model will determine the semantic relationship(s) among the content being analyzed, which results in a high degree of accuracy.

The Generative AI Module (Explanation Layer) sits at the center of the system but is the final piece of the overall architecture. The purpose of the Explanation Layer is to generate a readable human explanation of the AI-generated prediction and provide indications of how the prediction was determined (i.e., that the content is likely the result of misinformation, the technique used to perpetuate the misinformation, etc.). Thus, it will provide examples of common techniques used in misinformation (e.g., exaggeration; emotional manipulation; unsupported claims) to improve user confidence in the system and allow them to better understand the methodology used to determine the classification of the content.

Also, as the Prediction Layer is generating the machine-generated classification of the content, the External Knowledge Source Module is obtaining information from outside sources such as fact-checking APIs, newspaper publishers, etc., and correlating the machine-generated predictions against the actual verified real-world information. This will increase the credibility of the output since the prediction will be validated by both external data sources and machine-generated data.

Once all the data has been processed through each of the layers, the results will be delivered to the user through a user interface/web application or a browser extension as part of a user-friendly output. Each output will contain a label for the classification, the human explanation for the classification, external support for the classification, and tips on how to better identify misinformation.

Overall, the system's layered architecture provides a seamless flow of information from input to output while incorporating machine learning, Generative AI, and validated data from outside sources to develop a complete and user-friendly system for detecting misinformation.

IV. RESULT

V. CONCLUSION

The suggested AI-powered misinformation detection system is an efficient approach to combatting the growing problem of falsified or misleading information online. This system utilises state-of-the-art Natural Language Processing techniques in conjunction with transformer-based models to accurately classify written/typed content as true, false and/or misleading. The primary benefit of this method when compared with existing detection approaches, is that it not only classifies content, but utilises Generative AI to produce a clear, human-readable explanation of how it arrived at its conclusion. This feature increases the overall transparency of the system and thereby builds greater trust among users than will be found in existing systems. To further increase the reliability and credibility of the system, the output from the proposed AI system can be enhanced by sourcing prediction evidence from external knowledge sources.

Examples of such external knowledge sources include fact checking APIs. Together with the systems implementation method (as both a web application and web browser extension) ensures that it will be available to a diverse group of users, including students, researchers, and the general public; therefore providing real-time functionality.At this point in time, the functionality of the proposed AI system primarily focuses on detecting text-based misinformation; therefore it has limitations with respect to detecting multimedia and multiple language content types.

However, it provides an excellent foundation upon which future enhancements to the functional capabilities of the system may be developed. In conclusion, the proposed project will help to create a body of evidence that promotes digital literacy, responsible sharing of information, and informed decision making. By integrating detection, explanation, and education; the AI misinformation detection system offers a holistic and scalable solution to addressing the issue of misinformation in today's rapidly changed digital world.

Acknowledgments

We are pleased to convey our deep appreciation to our internal guide, P.Bharathi,Asst Prof. Dept. of IT, GRIET for her valuable encouragement, constructive comments and full support to complete our paper. She works as an Assistant Professor in the Department of Information Technology at Gokaraju Rangaraju Institute of Engineering and Technology (GRIET), Hyderabad, India. She have an academic experience of 15 years in teaching .With a strong academic background and a passion for research, she specializes in Machine Learning and Deep Learning. She has published research papers in reputed international journals and conferences. Her research interests include Deep Learning.

REFERENCES

1. OUEDRAOGO, N. (2020). SOCIAL MEDIA LITERACY IN CRISIS CONTEXT: FAKE NEWS CONSUMPTION DURING COVID-19 LOCKDOWN. AVAILABLE AT SSRN 3601466.
2. GUO, B., DING, Y., YAO, L., LIANG, Y., & YU, Z. (2020). THE FUTURE OF FALSE INFORMATION DETECTION ON SOCIAL MEDIA: NEW PERSPECTIVES AND TRENDS. ACM COMPUTING SURVEYS (CSUR), 53(4), 1-36.
3. MOLINA, M. D., SUNDAR, S. S., LE, T., & LEE, D. (2021). "FAKE NEWS" IS NOT SIMPLY FALSE INFORMATION: A CONCEPT EXPLICATION AND TAXONOMY OF ONLINE CONTENT. AMERICAN BEHAVIORAL SCIENTIST, 65(2), 180-212.
4. MISHRA, S., SHUKLA, P., & AGARWAL, R. (2022). ANALYZING MACHINE LEARNING ENABLED FAKE NEWS DETECTION TECHNIQUES FOR DIVERSIFIED DATASETS. WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, 2022(1), 1575365.
5. VERSTRAETE, M., BAMBAUER, J. R., & BAMBAUER, D. E. (2022). IDENTIFYING AND COUNTERING FAKE NEWS. HASTINGS LJ, 73, 821.
6. HOWARD, P. N. (2020). LIE MACHINES: HOW TO SAVE DEMOCRACY FROM TROLL ARMIES, DECEITFUL ROBOTS, JUNK NEWS OPERATIONS, AND POLITICAL OPERATIVES. YALE UNIVERSITY PRESS.
7. AHMED, K., KHAN, M. A., HAQ, I., AL MAZROA, A., SYAM, M. S., INNAB, N., ... & ALKAHTANI, H. K. (2024). SOCIAL MEDIA'S DARK SECRETS: A PROPAGATION, LEXICAL AND PSYCHOLINGUISTIC ORIENTED DEEP LEARNING APPROACH FOR FAKE NEWS PROLIFERATION. EXPERT SYSTEMS WITH APPLICATIONS, 255, 124650.
8. MAYORGA, M. W., HESTER, E. B., HELSEL, E., IVANOV, B., SELLNOW, T. L., SLOVIC, P., ... & FRAKES, D. (2020). ENHANCING PUBLIC RESISTANCE TO "FAKE NEWS" A REVIEW OF THE PROBLEM AND STRATEGIC SOLUTIONS. THE HANDBOOK OF APPLIED COMMUNICATION RESEARCH, 197-212.
9. SHU, K., MAHUDESWARAN, D., WANG, S., LEE, D., & LIU, H. (2020). FAKENEWSNET: A DATA REPOSITORY WITH NEWS CONTENT, SOCIAL CONTEXT, AND SPATIOTEMPORAL INFORMATION FOR STUDYING FAKE NEWS ON SOCIAL MEDIA. BIG DATA, 8(3), 171-188.
10. AHMAD, A., AZZEH, M., ALNAGI, E., ABU AL-HAIJA, Q., HALABI, D., AREF, A., & ABUHOUR, Y. (2024). HATE SPEECH DETECTION IN THE ARABIC LANGUAGE: CORPUS DESIGN, CONSTRUCTION, AND EVALUATION. FRONTIERS IN ARTIFICIAL INTELLIGENCE, 7, 1345445.
11. Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., ... & Sharma, R. (2022). A taxonomy of fake news classification techniques: Survey and implementation aspects. IEEE Access, 10, 30367-30394.
12. Amer, E., Kwak, K. S., & El-Sappagh, S. (2022). Context-based fake news detection model relying on deep learning models. Electronics, 11(8), 1255.
13. Piya, F. L., Karim, R., & Arefin, M. S. (2022). BDFN: a bilingual model to detect online fake news using machine learning technique. In Soft Computing for Security Applications: Proceedings of ICSCS 2021 (pp. 799-816). Springer Singapore.
14. Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032.

15. Shaikh, J., & Patil, R. (2020, December). Fake news detection using machine learning. In 2020 IEEE International symposium on sustainable energy, signal processing and cyber security (iSSSC) (pp. 1-5). IEEE.
16. Sharma, U., Saran, S., & Patil, S. M. (2020). Fake news detection using machine learning algorithms. International Journal of creative research thoughts (IJCRT), 8(6), 509-518.
17. Asha, J., & Meenakowshalya, A. (2021). Fake news detection using ngram analysis and machine learning algorithms. Journal of Mobile Computing, Communications & Mobile Networks, 8(1), 33-43p.
18. Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.
19. Klusowski, J. M., & Tian, P. M. (2024). Large scale prediction with decision trees. Journal of the American Statistical Association, 119(545), 525-537.
20. Schidler, A., & Szeider, S. (2024). SAT-based decision tree learning for large data sets. Journal of Artificial Intelligence Research, 80, 875- 918.