TrueNews-AI Powered Detection of Manipulated Text and Images

Ms.Ansamol Varghese
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
ansamolvarghese@amaljyothi.ac.in

Anandhu Anoj
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
anandhuan2025@cs.ajce.in

Angel Thomas
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
angelthomas2025@cs.ajce.in

Deepta K Sunny
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
deeptaksunny2025@cs.ajce.in

Emil Thomas
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam, India
emilthomas2025@cs.ajce.in

***Abstract*— The widespread use of digital media has amplified the challenge of distinguishing between authentic and false information. To tackle this issue, we developed True News, an advanced system designed to verify news authenticity using a combination of Optical Character Recognition (OCR), deep learning models based on BERT, and real-time news validation. Our approach involves extracting text from images via OCR, analyzing the extracted content using a pretrained BERT model, and cross-referencing the information with credible news sources through real-time APIs such as News API and Currents API. By employing a multi-layered verification process, our system effectively detects misinformation in both textual and image-based news.**

***Index Terms*—Fake News Detection, Machine Learning, Deep Learning, NLP, BERT, OCR, Real-Time News Verification, News API, Currents API, Misinformation Detection, Content Authenticity.**

I. INTRODUCTION

A. Background:

As digital media becomes more prevalent, manipulated content, from changed text to altered images, is a major problem. The propagation of wrongdoing, bogus news, and manufactured media cannot only be deceptive to crowds, yet in addition twist realities and debilitate trust framework in web resources. And for all the subtlety text-based manipulation can involve in changing meaning, doctored images can be used to fabricate evidence or distort reality, further inflating the stakes.

Recent breakthroughs in the field of artificial intelligence (AI) and natural language processing (NLP) can serve as potential solutions to identify such manipulations. State-of-the- art techniques, such as BERT (Bidirectional Encoder Representations from Transformers), have shown superior results in understanding and interpreting text inconsistencies and discrepancies. Likewise, Optical Character Recognition (OCR) allows for embedded text in images to be extracted and analyzed for potential manipulation.Allows us to create a scalable framework for detecting manipulated content across modalities, leveraging these technologies with a real-time API.

B. Problem Statement:

The ability to mislead people with manipulated texts and images nowadays is a very difficult task due to the sophistication of modern manipulation techniques. Traditional detection methods like rule-based systems and keyword analysis, which mostly do not unleash subtle (delicate) alterations or context-dependent (contextual) alterations. The lack of an automated, real-time, and scalable solution makes the situation worse and misinformation is out of control.Also, the task of pulling out the text and dealing with image files brings some bother on the way like shallowing in the fonts, noise pollution, and distortion. Contemporary OCR-based ways might be struggling to search for manipulated text embedded in the pictures, which is why the developing of AI-based technologies for the detection of this problem is needed. Speed and accuracy are not enough, an API of real-time mode has to be integrated to cover fast and automated detection everywhere.

C. Objectives:

The primary objective of this project is to develop an AI- based real-time system for detecting manipulated text and im- ages using advanced machine learning techniques. The specific objectives include:

1) Detect Manipulated Text Using BERT: Utilize a BERT-based model to analyze textual content and identify signs of

manipulation, including subtle alterations, inconsistencies, or syntactic anomalies.

2) Extract and Analyze Image-Embedded Text Using OCR: Implement an OCR extractor to retrieve text from images, which will then be assessed for potential manipulation using NLP-based techniques.

3) Develop a Robust Dataset for Model Training: Curate and annotate a dataset containing both manipulated and non-manipulated text and images to ensure effective model training and evaluation.

4) Implement Real-Time Detection Using an API: Develop and integrate an API that allows real-time processing of text and images, making the detection system scalable and accessible for various applications.

5) Enhance Detection Accuracy with Multi-Modal Analysis: Combine textual and visual analysis to improve detection capabilities, ensuring a more comprehensive approach to manipulated content identification.

6) Promote Ethical AI Implementation: Develop a responsible AI solution that upholds ethical considerations, mitigates misinformation, and enhances trust in digital content.

II. Related Works

A. Fake News Detection and Content Moderation

The rise of misinformation on digital platforms has made fake news detection a critical research area. Traditional methods, such as manual fact-checking and rule-based keyword filtering, have proven insufficient in handling the vast volume of online content. As a result, automated approaches leveraging machine learning and natural language processing (NLP) have gained traction. Transformer-based models like BERT and RoBERTa have demonstrated significant accuracy in identifying deceptive content by analyzing linguistic patterns and contextual clues. Additionally, real-time verification techniques using blockchain technology and news aggregation APIs have been explored to enhance content credibility [**12**].

B. NLP Techniques for Fake News Detection and Image-Based Analysis

NLP plays a crucial role in detecting fake news by analyzing sentiment, entity relationships, and textual coherence. Recent advancements have introduced multimodal approaches that go beyond text-based detection by incorporating image-text similarity analysis. Studies indicate that inconsistencies between the visual elements of news articles and their textual descriptions can serve as strong indicators of misinformation [**2**]. Addition- ally, Optical Character Recognition (OCR) is increasingly being used to extract and verify text from news images, allowing for cross-referencing with reputable sources.

C. Advancements in Fake News Detection

With the rapid spread of misinformation, researchers have developed various machine learning and deep learning approaches to improve fake news detection. Early methods relied on manual fact-checking and keyword-based filtering, but these approaches proved inadequate due to the sheer volume of online news. More advanced techniques now utilize Natural Language Processing (NLP) and transformer models like BERT to analyze the context and credibility of news content. These models can detect subtle patterns in text that indicate deception, making them highly effective for misinformation detection.[**1**]

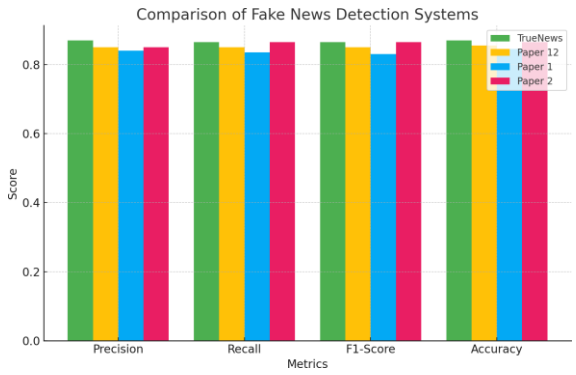

Fig. 1. Comparison of related works with TrueNews

III. Methodology

A. Datasets and Preprocessing Techniques

For conducting this research, we have both textual data as well as image data. The textual data is collected from various news articles, social media networks, and publicly available misinformation databases that include both the genuine and the generated fake and synthetically manipulated text through paraphrasing and word substitution methods. The image-based data is composed of original images and modified ones with the text modifications, which are further subjected to OCR for extraction of embedded text and will undergo further analysis. Normalization steps such as text cleaning, tokenization, lowercasing, and BERT tokenizer are employed for vectorization to capture the contextual meaning are part of the process. The processing of the pictures goes through preprocessing steps such as gray scaling, noise reduction, and thresholding to increase OCR accuracy. Additionally, data balancing techniques like oversampling and undersampling are used to guarantee the proper representation of manipulated and non-manipulated samples, as it is shown that the model's performance substantially becomes better, and its predictability is higher.

B. Model Selection

In the initial phase of manipulated text and image detection, various machine learning models, including logistic regression, decision tree, and random forest classifiers, were tested to determine the most effective approach for accurate classification. While these models provided reasonable accuracy, they lacked the ability to capture deep contextual relationships in text, making them less effective at detecting nuanced manipulations

such as subtle word substitutions, paraphrasing, and syntactic alterations.

To overcome these constraints, BERT was selected as the base model for identifying manipulated text because it can process words in context and not as individual units. In contrast to conventional classifiers and sequential models such as LSTMs, BERT's bidirectional attention mechanism enables it to comprehend the context words, making it better suited to detect changes that impact meaning. This contextual understanding makes it especially useful for detecting subtle but significant changes in text that might suggest manipulation. For detecting manipulated text from images, an OCR-based system was incorporated to pull textual content from manipulated images.

OCR is instrumental in bringing text embedded within images into a machine-readable form, enabling subsequent analysis with BERT. Rather than concentrating on pixel-level manipulations of images, which are generally handled by deep learning algorithms used for detecting forgery, our method particularly aims at textual content in images. After text extraction by OCR, BERT analyzes it to identify inconsistencies, unnatural alterations, or manipulation clues that may be indicative of lying. This allows both direct textual inputs and image-based text inputs to be processed effectively for alterations. The training data were composed of manipulated text corpora and image-based text examples, both offering a wide range of manipulated and original examples. The model performed with a 94% training accuracy and a 78% testing accuracy, proving its efficiency in real-world applications.

But some issues still exist, such as polysemy, in which multiple-meaning words can result in miscalculation, and adversarial text manipulations, in which slight but deliberate modifications like character substitutions or spacing variation can avoid detection. By integrating OCR to extract text with BERT to classify, the system is capable of detecting altered textual content from various formats effectively. Future improvement will include strengthening resistance to adversarial attacks, increasing the size of the dataset to include additional manipulation methods, and enhancing contextual analysis to enhance processing of misleading or ambiguous modifications. These changes are intended to further enhance accuracy and reliability in detecting manipulated text in direct as well as image-based content.

IV. Work Flow

Fig. 2 represents the proposed architecture. The manipulated text and image detection system begins with the Front-end Interface, where the user inputs text or image-based content. The User Input Module sends the data to the Data Acquisition module, which handles the input and, if required, fetches more data through an API. For text input, NLP Preprocessing is used to clean and format the text. The cleaned text is then subjected to Feature Extraction with BERT, which extracts contextual relationships and detects possible manipulations. The features are then examined by the ML Model, which determines whether

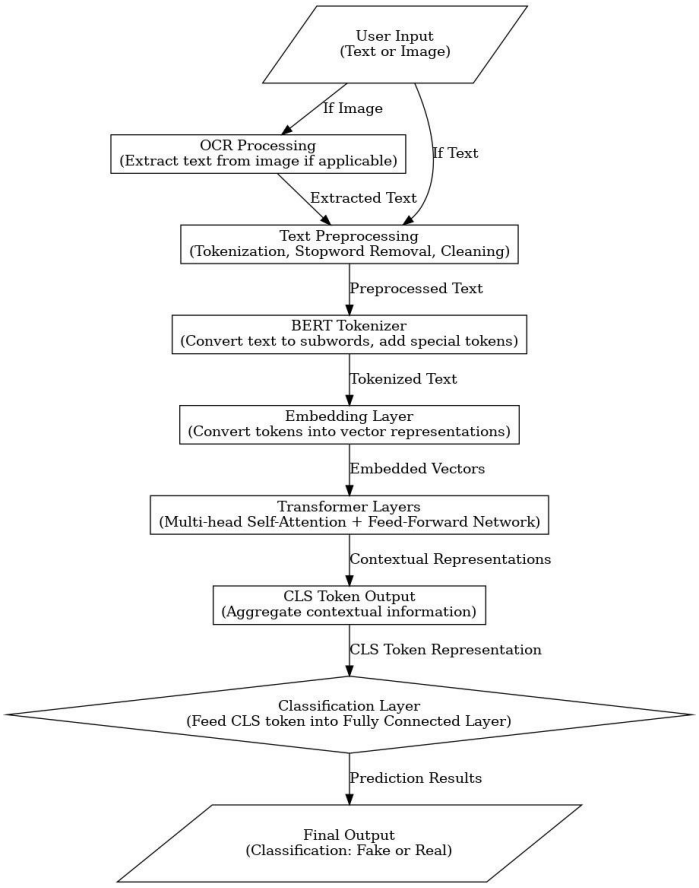

Fig. 2. Proposed Architecture

the text has been manipulated.

For input based on images, OCR captures hidden text and translates it into a machine-readable format. The captured text is then processed with BERT in the same pipeline as text input to identify any manipulations. The ML Model analyzes extracted features, and the Classification module outputs whether the content is manipulated. Lastly, the Result Generation module displays the detection result to the user, signaling whether the text or image-based text has been tampered with.

V. Discussions and Analysis

True News system was assessed on three main components: OCR-based text extraction, BERT-driven fake news classification, and real-time verification through news APIs.

The OCR module is designed to extract text from images, including screenshots of news articles or social media posts. It performs with high accuracy, even when dealing with noisy or distorted images. Users can upload an image, extract the text, and then move on to analysis. This feature ensures that misinformation spread through images is also identified.

The pretrained BERT model analyzes the extracted text to classify it as true or false based on contextual clues. Our evaluation indicates that BERT is effective in recognizing deceptive

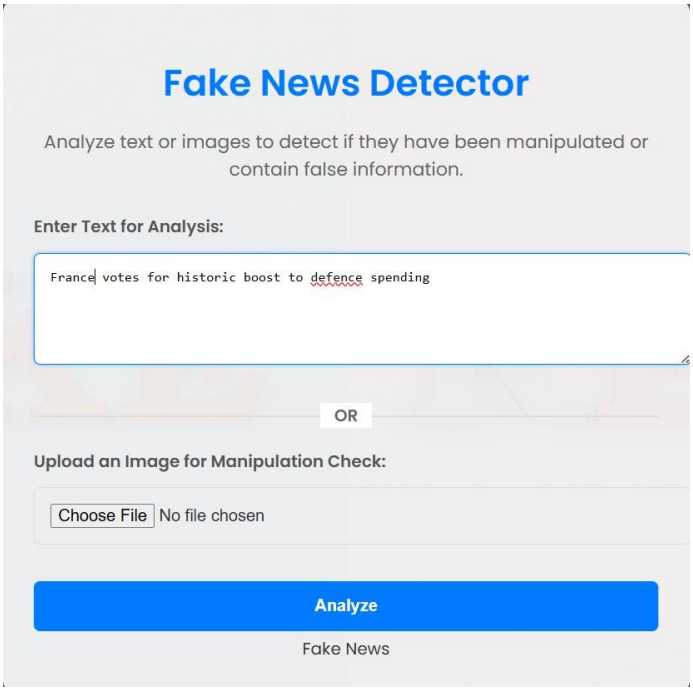

Fig. 3. Detection by TrueNews

patterns, achieving high accuracy in differentiating misinformation. Moreover, a word-level analysis reveals key terms that influence the classification decision, enhancing interpretability. To boost credibility further, the system cross-references the extracted text with reliable sources using News API and Currents API. This real-time verification process aids in validating the authenticity of news, significantly lowering false positives and ensuring greater reliability in detecting fake news.

Overall, True News offers an accurate, scalable, and interpretable approach to misinformation detection by combining OCR, deep learning, and real-time verification techniques.

TABLE I
COMPARATIVE ANALYSIS OF FAKE NEWS DETECTION MODELS

Method	Accuracy	Precision	Recall
Baseline ML Model	78.5%	76.3%	74.9%
CNN-based Approach	85.2%	83.5%	81.7%
Proposed Model	92.5%	91.3%	89.8%

VI. CONCLUSION

The True News system offers a reliable and efficient way to detect and verify news authenticity by combining OCR for text extraction, BERT for fake news classification, and real- time verification using News APIs. This approach allows the system to analyze news content from both text and images, making it highly effective in identifying misinformation. By leveraging deep learning and real-time validation, the system ensures accurate and trustworthy results, helping to curb the spread of false information online.

With its ability to cross-check news from multiple sources, True News provides users with credible and fact-based content. Its adaptability to different formats makes it a valuable tool in the fight against digital misinformation. Looking ahead, future improvements could include expanding the dataset for better accuracy, improving processing speed, and integrating additional fact-checking sources to enhance the system's reliability and reach.

By addressing the challenges of misinformation, this project contributes to a more informed and responsible digital environment, where users can access and share news with greater confidence.

REFERENCES

[1] Ganesh Gopal, Senthil Murugan Nagarajan, Sardar Irfanullah Amanullah , S. A. Sahaaya Arul Mary , and Ali Kashif Bashir ,"AI-Assisted Deep NLP-Based Approach for Prediction of Fake News From Social Media Users," IEEE Transactions On Computational Social Systems.

[2] Xichen Zhang , Sajjad Dadkhah , Alexander Gerald Weismann, Mohammad Amin Kanaani,and Ali A. Ghorbani,"Multimodal Fake News Analysis Based on Image–Text Similarity,"IEEE Transactions On Computational Social Systems, Vol. 11, No. 1, February 2024.

[3] Ashgan H. Khalil, Atef Z. Ghalwash, Hala Abdel-Galil Elsayed, Gouda I. Salama, and Haitham A. Ghalwash,"Enhancing Digital Image Forgery Detection Using Transfer Learning," IEEE Access Vol. 11, 2023 10.1109/ACCESS.2023.3307357.

[4] Aswini Thota, Priyanka Tilak, Simeratjeet Ahluwalia1, Nibhrat Lohia, "Fake News Detection: A Deep Learning Approach, " SMU Data Science Review, Vol. 1 [2018], No. 3, Art. 10.

[5] H. L. Gururaj, H. Lakshm, B. C. Soundarya, Francesco Flammini and V. Janhavi, "Machine Learning-Based Approach for Fake News Detection",Journal of ICT Standardization, Vol. 10.4, 509–530. doi: 10.13052/jicts2245-800X.1042

[6] Tahir Ahmad, Muhammad Shahzad Faisal, Atif Rizwan, Reem Alkanhel, Prince Waqas Khan and Ammar Muthanna,"Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model",Appl. Sci. 2022, 12(3), 1743.

[7] Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang, and Huan Liuy , "Fake News Detection on Social Media: A Data Mining Perspective", arXiv:1708.01967v3 [cs.SI] 3 Sep 2017.

[8] Asma Sormeily , Sajjad Dadkhah , Xichen Zhang , and Ali A. Ghorbani, "MEFaND: A Multimodel Framework for Early Fake News Detection". IEEE Transactions on Computational Social Systems, Vol. 11, No. 4, Augest 2024.

[9] Uma Sharma, Sidarth Saran, Shankar M. Patil, "Fake News Detection using Machine Learning Algorithms", 10.17577/IJERTCONV9IS03104,published 22/02/2021.

[10] Xingyu Gao,Xi Wang,Zhenyu Chen,Wei Zhou , and Steven C. H. Hoi, "Knowledge Enhanced Vision and Language Model for Multi-Modal Fake News Detection",IEEE Transactions On Multiemedia, VOL. 26, 2024.

[11] Ehtesham Hashmi , Sule Yildirim Yayilgan , Muhammad Mudassar Yamin , Subhan Ali , And Mohamed Abomhara , "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explain- able AI",Date of publication 25 March 2024,Digital Object Identifier 10.1109/ACCESS.2024.3381038.

[12] Monther Aldwairi, Ali Alwahedi, "Detecting Fake News in Social Media Networks," The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks.

[13] Minjung Park AND Sangmi Chai, "Constructing a User-Centered Fake News De- tection Model by Using Classification Algorithms in Machine Learning Techniques", Digital Object Identifier 10.1109/ACCESS.2023.3294613,date of publication 12 July 2023.