WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

TruthLens: Multimodal Fake Content Detector

K Bhargavi [1], Nitya [2], Pragnya [3], Sri Amar [4]

Dept. Of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalabuargi, *India*
bhargavikarankote@gmail.com [1],nityakulkarni11@gmail.com[2] ,pragnyakottargi@gmail.com[3], amarchinti@pdaengg.com [4]

ABSTRACT

In today's digital landscape, the rapid and widespread dissemination of false information has emerged as acritical global challenge. The issue is further intensified by the increasing sophistication of artificially generated text, manipulated images, and coordinated misinformation campaigns that operate across multiple languages and platforms. In response to this growing threat, this project proposes a comprehensive Multi-Modal Fake Content Detection System designed to evaluate and verify the authenticity of digital content shared online. The proposed system is built around three integrated modules. The first focuses on analyzing textual content across various languages, identifying misleading or deceptive information through advanced linguistic and contextual analysis. The second module examines visual content, detecting signs of manipulation or fabrication within images. The third component cross-references textual data specifically to detect fake or misleading narratives, enhancing detection through deeper semantic understanding. By combining insights from both textual and visual sources, this multi-modal approach strengthens the system's ability to detect inconsistencies and anomalies that may indicate fake content. The synergy between these modules significantly boosts overall accuracy and reliability. This detection system can be instrumental in various sectors, including journalism, social media oversight, and digital forensics. It offers a powerful tool for identifying and curbing the spread of misinformation, ultimately contributing to a more trustworthy and informed digital environment.

INTRODUCTION

The widespread circulation of false news articles, manipulated images, and misleading posts across multiple languages poses a significant threat to democratic processes, fuels panic during emergencies, and diminishes public trust. In the digital age, information disseminates rapidly through online platforms, reaching millions of users within seconds. While this revolutionizes communication, it simultaneously exacerbates the problem of fake content proliferation. Traditional fake content detection systems have predominantly focused on single content formats—primarily text or images—and monolingual settings, which are increasingly inadequate. Contemporary fake content is often multimodal and multilingual, presenting significant challenges to existing detection techniques. Addressing these challenges requires systems capable of processing diverse languages and transforming them into a unified format for effective analysis. Recent advances in multilingual neural machine translation models, combined with natural language processing and classical machine learning classifiers, provide promising avenues for detecting fake news across languages. This approach leverages automated language detection, translation, and text preprocessing techniques to improve the reliability and applicability of fake news detection in a global, multilingual context.

1. Text Content Fake Detection Module

Text remains the most common medium through which fake news is propagated, especially on platforms like Twitter, Facebook, and news websites. Textual misinformation can range from fabricated stories and conspiracy theories to misleading headlines and biased reporting. Detecting such content requires not only syntactic analysis but also an understanding of semantic meaning, context, and intent.

In this module, we combine traditional machine learning and deep learning techniques for fake news detection. Text data is first preprocessed using NLP steps like tokenization, stop word removal, and lemmatization. Features are extracted using TF-IDF vectorization, and the data is balanced and split for training. A Keras-based deep learning model with dense layers and dropout is used for classification, alongside a Multinomial Naive Bayes model for comparison. The system is evaluated using accuracy, precision, recall, confusion matrix, and ROC-AUC to ensure reliable performance.

2. Multilingual Fake Content Detection Module

In a globalized world, misinformation transcends linguistic boundaries. Content originally created in one language can quickly spread across regions, translated (often inaccurately or intentionally altered) into many others. Therefore, a robust fake content detection system must be capable of handling multilingual input and understanding cultural and linguistic nuances.

In this module, we combine traditional machine learning and deep learning techniques for fake news detection to enhance accuracy and multilingual capability. Traditional models like Naïve Bayes are employed alongside TF-IDF vectorization to classify news content based on textual features. To address the multilingual nature of online misinformation, we incorporate Facebook's M2M100 translation model, which identifies the source language

and translates it to English without relying on English-centric assumptions. This allows the system to process and analyze news articles written in various languages using a unified pipeline. The integration of language detection, machine translation, and classical classification provides a robust framework for identifying fake content across linguistic boundaries.

3. Image Fake Detection Module

Visual content plays a powerful role in shaping public perception. In recent years, manipulated images and deepfakes have emerged as a new form of disinformation. Images can be altered to change their context, taken out of context, or entirely fabricated using AI tools like GANs (Generative Adversarial Networks). Relying solely on human judgment to detect these manipulations is no longer viable, especially at scale.

In this module, we address fake image detection using deep learning-based computer vision techniques. We utilize a Vision Transformer (ViT) model, fine-tuned specifically for deepfake detection. The image input is first processed using a pretrained AutoImageProcessor, and then passed through the AutoModelForImageClassification from the Hugging Face Transformers library. The model, sourced from 'ashish-001/deepfake-detection-using-ViT', is capable of identifying subtle artifacts and manipulations in images that are often associated with synthetic or tampered content. This approach ensures robust detection by leveraging advanced attention mechanisms in ViT architecture.

RELATED WORK

Ye Zhu, Yunan Wang, Zitong, 2025, [1], introduces the MFND dataset and proposes a Shallow-Deep Multitask Learning (SDML) model that fuses image and text features for fake news detection and localization. Eunjee Choi, Junhyun Ahn, XinYu Piao, Jong- Kook Kim,2025, [2], Proposes the CroMe model, which utilizes Cross-Modal Tri-Transformer and metric learning to enhance multimodal fake news detection. Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, Dong Wang, 2025, [3], Introduces MIMoE-FND, a hierarchical Mixture-of-Experts framework that models modality interactions for improved fake news detection.

Tianlin Zhang, En Yu, Yi Shao, Shuai Li, Sujuan Hou, Jiande Sun,2025,[4], proposes MIAN, a framework that leverages inverse attention mechanisms to highlight conflicting patterns in multimodal data for fake news detection. "FakeBERT: Fake News Detection in Social Media with a BERTbased Deep Learning Approach",Rohit Kumar Kaliyar,Anurag Goswami,Pratik Narang[5] FakeBERT combines BERT's deep contextual embeddings with parallel 1D-CNN layers to effectively capture semantic patterns in news text. It achieves 98.90% accuracy on a real-world dataset, outperforming traditional models in detecting fake news with low error rates.

PROPOSED METHODOLOGY

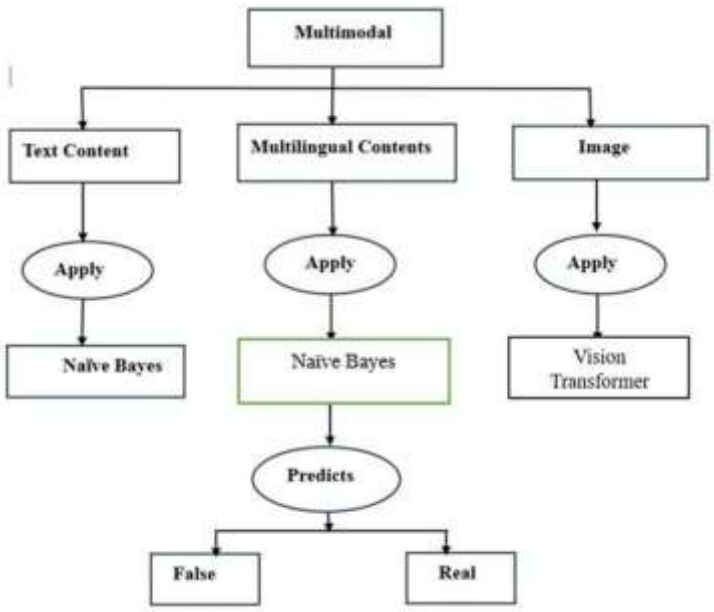

Fig 1: System Architecture

The proposed framework presents a multimodal fake content detection architecture capable of analyzing and classifying diverse data types—namely textual, multilingual, and visual content—to determine their authenticity. This system is strategically designed to assign specialized models to each modality, thereby enhancing accuracy and robustness in distinguishing between real and manipulated inputs, which is critical in combating the widespread dissemination of misinformation and synthetic media.

For monolingual text-based inputs, such as news articles or social media posts in a single language, the system initiates classification using the Naïve Bayes algorithm. The Naïve Bayes classifier is a probabilistic model grounded in Bayes' Theorem, which assumes conditional independence among input

features given the target class. Despite its simplicity, it proves to be highly effective for natural language processing tasks due to the sparse and high-dimensional nature of text data. It estimates posterior class probabilities using prior knowledge and observed word frequencies, offering fast and interpretable classification for distinguishing fake and real textual information.

For content presented in languages other than English, the system routes the input through a dedicated multilingual processing pipeline that supports various linguistic structures and scripts. This input is either directly tokenized or translated and then classified using the same Naïve Bayes approach. This ensures that the detection system remains language-agnostic and effective across diverse geopolitical and cultural contexts where misinformation may be presented in non-English languages.

In the case of visual content, particularly images suspected to be altered or artificially generated the system implements a specialized image classification module.

Face Detection using Haar Cascade:

Prior to deepfake analysis, facial regions are identified using the Haar Cascade classifier, which is an efficient object detection method based on simple rectangular features. This step isolates relevant facial areas, significantly reducing computational complexity and focusing the model on regions most likely to exhibit manipulation artifacts.

Vision Transformer (ViT) :

To classify the extracted facial regions, the system employs the Vision Transformer (ViT), a cutting-edge deep learning architecture adapted from the transformer models used in NLP. ViT transforms images into a sequence of fixed-size patches and applies self-attention mechanisms to capture long-range dependencies and spatial relationships across the entire image. Its ability to model global contextual information makes it well-suited for detecting subtle, spatially dispersed tampering found in fake or synthesized images.

Following the individual modality-specific analyses, outputs from the textual, multilingual, and visual pipelines are aggregated into a unified classification stage that produces a binary prediction: either Real or Fake. The integration of traditional statistical models like Naïve Bayes with advanced architectures such as ViT, along with preprocessing enhancements like Haar Cascade, results in a robust and scalable solution for multimodal fake content detection in diverse digital ecosystems.

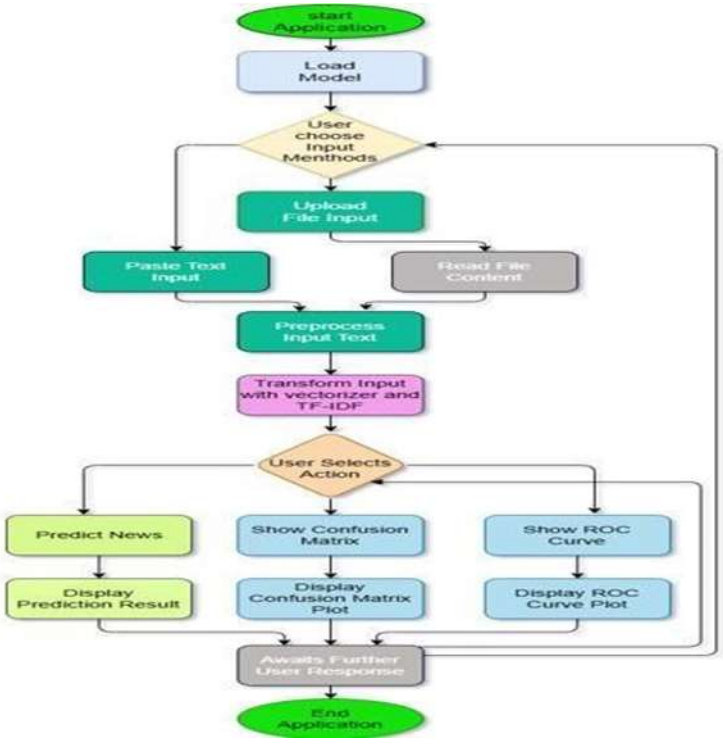

Fig 2 : Text detection flow diagram

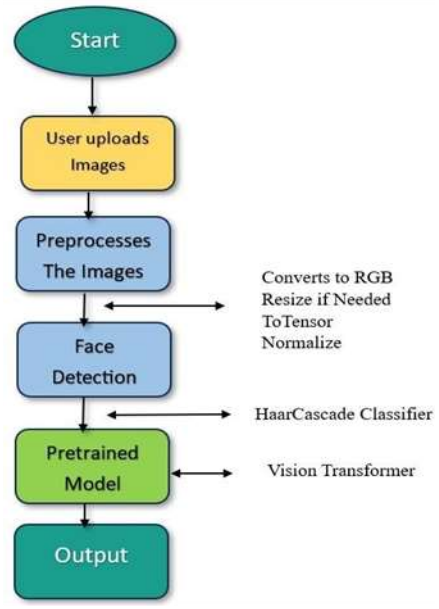

Fig 3 : Image detection flow diagram

RESULT AND DISCUSSION

The proposed Multi-Modal Fake Content Detection System achieved high accuracy in detecting fake content across text, multilingual text, and image modalities. The Naïve Bayes classifier effectively identified fake textual content with reliable precision and recall. The multilingual module, using M2M100 for translation and TF-IDF with Naïve Bayes, maintained performance across diverse languages. The ViTbased image module successfully

detected manipulated visuals, especially deepfakes, using facial region focus via Haar Cascade. The system's modular integration led to enhanced detection accuracy and robustness.

Fig 4 : Text Content Detection

Fig 5:Image detection

CONCLUSION

This project presents a comprehensive and scalable approach to combating misinformation by leveraging a multimodal detection system. The fusion of text, multilingual, and image analysis provides a holistic method for evaluating digital content authenticity. Naïve Bayes proved efficient for textual classification due to its simplicity and effectiveness in high-dimensional data. The inclusion of multilingual translation pipelines ensures global applicability. The Haar Cascade pre-processing combined with Vision Transformer architecture enhances visual detection accuracy, especially for deepfake identification. By aggregating outputs from all modules, the system delivers reliable binary predictions, minimizing false positives and negatives. Its cross- platform and cross-language capabilities make it suitable for real-world deployment in journalism, content moderation, and digital forensics. This system not only improves misinformation detection but also fosters a safer, more trustworthy online environment.