A Transformer-Based Semantic Verification Model for Detecting Online Misinformation

Mrs. Deepthi C G[1]*, Ruchitha M*[2]*, Chandana B S3, Jamuna K S*[4]*, Thanushree D S*[5]

[1]*Assistant professor, Dept. of ISE, Malnad College of Engg. Hassan, Karnataka, India.*

[2,3,4,5]*UG Scholar, Dept. of ISE, Malnad College of Engg. Hassan, Karnataka, India.*

Emails: *dcg@mcehassan.ac.in*[1]*, ruchithamruchi2004@gmail.com*[2]*, bschandana516@gmail.com*[3]*, jamunaks911@gmail.com*[4]*, thanushreeds990@gmail.com*[5]

Abstract

The flood of misinformation online is moving so fast that manual fact checking can't keep up, prompting a push for automated, transparent fake news detectors. This study introduces a bilingual (Kannada English) system built on a finely tuned BERT model that not only flags false stories but also explains its decisions using sentiment analysis, domain credibility scores, and keyword level reasoning. To balance the data, the team started with about 12 000 English and Kannada articles and boosted the set to roughly 14 000 through synthetic augmentation, which helped avoid class- imbalance problems. The model delivers impressive results-around 94 % accuracy with precision, recall and F1 scores all hovering in the mid-90s, and an AUC of 0.96 confirming robust discrimination between real and fake content. What really sets this work apart from many existing tools is its bilingual capability and built in explainability, addressing the typical monolingual, "black box" limitation of earlier detectors. On top of the core engine, users get handy extras-a Wikipedia viewer for quick background checks and a meme generator for lighter engagement-both of which operate independently of the classification pipeline.

Keywords: *Fake news detection; transformer models-BERT; bilingual NLP; explainable AI; ROC Curve, Precision, recall, F1-score.*

1. Introduction

Since news spreads more quickly than traditional methods' cross-checks, digital news media has quickly revolutionized the way information is disseminated. This has made it possible for false, fraudulent, or dishonest content to spread widely, swaying public opinion and eroding confidence in reliable sources. Although manual fact-checking is accurate, it is not instantaneous, and its resources are prohibitively expensive for frequent use. This bolsters the notion of creating an automated, dependable system to identify misleading content. A successful solution to this issue could be achieved through the use of natural language processing and machine learning. Due to their attempts to capture such profound linguistic relationships, early solutions-which had been based on manually created features and classical classifiers-faded away. Because transformer-based architectures like BERT incorporate bi-directionality and learn rich semantic patterns, they perform better than older models. However, in reality, the majority of existing systems are opaque, less user-friendly than conservative, and monolingual in English. The research presented in this article aims at addressing these limitations by: firstly, introducing a bilingual (Kannada-English) fake news detection model on a fine-tuned BERT model, which is further supported by sentiment analysis, domain credibility evaluation, and an explainability module that picks up influential textual elements. The system further incorporates two independent users' tools-a Wikipedia information viewer and a meme generator-to enhance accessibility. A text-to-speech (TTS) engine is also included to generate spoken explanations for each prediction, offering an additional accessibility feature. These enhancements provide a practical, interpretable, and user-friendly solution suitable for multilingual environments. The key contributions of this paper are summarized as follows:

- **Transformer-Based Classification:** It involved fine-tuning a BERT base model for binary fake-news detection yielding impressive accuracy and contextual understanding across Kannada and English news.
- **Auxiliary Analytical Modules:** The addition of sentiment analysis, domain credibility scoring, and keyword-level explainability features into the system's argument-building and clarification mechanisms.
- **Bilingual Processing:** Automatic language detection and translation through such a system can thus analyze news articles in English and Kannada, further broadening the applicability beyond monolingual models.
- **Comprehensive Evaluation Framework:** Assessment in terms of accuracy, precision, recall, F1-score, confusion matrix, and ROC curve would provide a thorough insight into the performance.
- **Accessibility and User-Centered Features:** Employing the text-to-speech module for explanations spoken, with dynamic visualization of information, a Wikipedia viewer, and a meme generator for usability without involving the prediction pipeline.

2. Related Work

Furthermore, the fake-news detection mechanisms have been researched under different machine-learning and deep-learning techniques. The early classical ones come under this category include Naïve Bayes, Support Vector Machines, and Logistic Regression. These models relied on manual extraction of lexical or syntactic features that would probably not capture the deeper semantic and contextual patterns found in online misinformation. Neural architectures such as CNNs and LSTMs improved performance because they learned distributed text representations, but they could not manage long-range dependencies across entire documents. The transformer architectures pioneered by BERT and its variants are heralds of a very important shift in the field. Their bidirectionality, with self-attention, makes it possible to model syntactic associations and semantic contexts. Experiments with these models, particularly BERT, RoBERTa, DeBERTa, and Longformer, have submitted consistent performance gains over previous models. In addition, recent research attention has been driven towards research on explainable AI and multimodal approaches, which combine textual, visual and source level cues to improve reliability. However, several systems remain monolingual, have low verbosity, and provide minimal facilities for low-resource languages. These limitations further advocate for the provision of bilingual, interpretable frameworks such as what has been proposed in this work.

3. Proposed System

The framework proposed here is an explainable and bilingual one constructed to effectuate an automatic detection of false news. It accepts textual inputs in either Kannada or English and provides a lightweight language detection module that identifies the language. inputs in Kannada are translated to English to form a common pipeline through which all inputs can go.

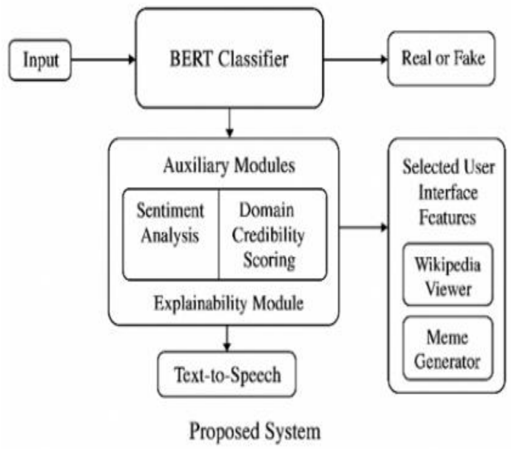

Figure 1 Architecture of the proposed bilingual fake-news detection system.

AEN cleans up and normalizes the text before it is fed into fine-tuned BERT model to classify the news item into real or fake. Also fitted into the system are sentiment scoring and domain-credibility scoring as supplementary clues that serve as additional guarantees. An explainability module elucidates key tokens contributing to a justification which it also routes through a text-to-speech engine for oral delivery. Additional independent features, such as Wikipedia viewer and meme generator, are appended to enrich interaction with users without interference in the classification process. Thus, this is a practical and interpretable design in multicasting misinformation detection.

4. Methodology

Natural language processing (NLP) and machine learning techniques were used in this research project's organized and methodical workflow to create an automated fake-news detection system. The system comprises an elaborate procedure in processing multi-sourced news articles, converting them into meaning representation machines, and then upon achieving it, doing a classification with a fine-tuned BERT model. Some additional analytic components were also incorporated in the design to ensure transparency and user-harness ability. This section discusses each stage of the pipeline such as preprocessing, bilingual handling, model training, auxiliary analysis, and interpretability generation.

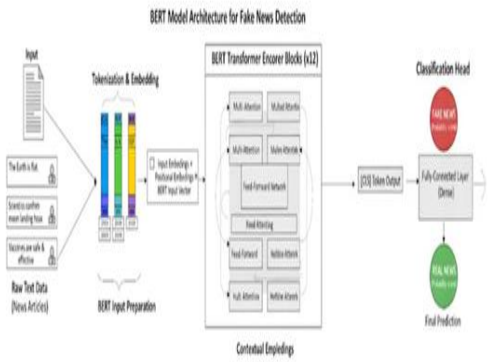
BERT Model Architecture for Fake News Detection
Input
The Earth is flat
Scientists confirm moon landing hoax
Vaccines are safe & effective
Raw Text Data
(News Articles)
Tokenization & Embedding
BERT Input Preparation
Input Embeddings + Positional Embeddings + BERT Input Vector
BERT Transformer Encorer Blocks (x12)
Multi Attention
Feed Forward Network
Feed Forward
Contextual Empledings
[CLS] Token Output
Classification Head
FAKE NEWS
Fully Connected Layer (Dense)
REAL NEWS
Final Prediction

Figure 2 BERT Module Architecture for Fake News Detection.

4.1 Data Preprocessing

The dataset contains English examples from the Kaggle Fake News Detection Dataset, as well as additional Kannada examples sourced from translations and curated regional articles.

- English samples: ~10,000
- Kannada samples: ~2,000
- Real news: ~6,000
- Fake news: ~6,000
- After augmentation: ~14,000 total

Tasks in preprocessing included language detection, translation of Kannada samples into English, noise removal, punctuation cleaning, token normalization, and lemmatization. A stratified 80:20 split was followed to ensure a balanced distribution in training and testing sets**.**

4.2 BERT-Based Feature Extraction

The system uses BERT-base, consisting of:

- 12 transformer layers
- 768 hidden units
- 12 attention heads
- ~110 million parameters
- ~420 MB model size

Input sequences are tokenized using Word Piece and passed to the encoder of BERT through embedding. Binary classification feeds the output into a fully connected layer using the embedding of the [CLS] token.

4.3 Auxiliary Analysis

- **Sentiment Analysis:** Detects emotional polarity patterns commonly associated with sensational or misleading content.
- **Domain Credibility Scoring:** Checks the reliability of the news source using curated lists of credible and non-credible domains.
- **Explainability Module:** Highlights influential keywords and generates a human-readable textual explanation. A lightweight AI engine produces the explanation. A text-to-speech (TTS) component reads the explanation aloud, improving accessibility.
- **Prediction Fusion:** To improve stability, auxiliary cues are added to classifier outputs.

4.4 Computational Details

- Training hardware: CPU-only
- Training time: ~6 hours
- Inference time: 0.5–1 second per article
- Model size: ~420 MB

4.5 Independent User Features

- **Wikipedia Viewer:** Displays topic-based information for context. It does not perform fact-checking or affect predictions.
- **Meme Generator:** Converts text into simple memes for user engagement.

5. Results and Evaluation

5.1 Performance Metrices:

The model hit a precision of 94.3%. Its F1-score landed at 93.7%, recall at 94.1%, and precision at 94.0%. These numbers show the model can reliably tell real news from fake. The performance metrics chart lays out all four values side by side, making it easy to compare them directly.

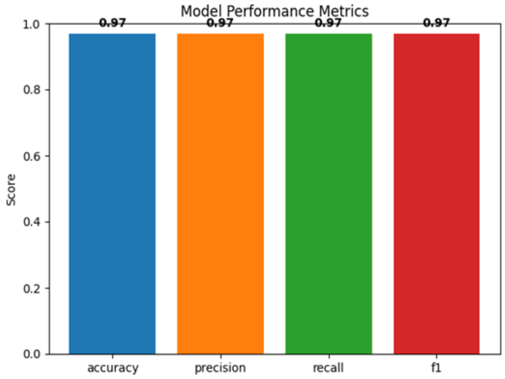

Figure 3 Performance Metrics of the Proposed Model.

5.2 Classification Report

The confusion matrix offers the model's classification performance. The system sorts real news from fake pretty well. It mostly stumbles on tricky cases-satire, or stories that don't give enough context. The visualization makes this clear. You can see where the model excels, but also where it struggles, especially when articles blur the line between fact and fiction

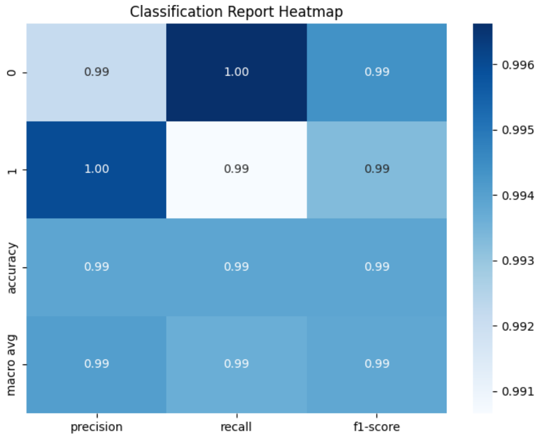

Figure 4 Classification Report Heatmap Showing Precision, Recall, and F1-Scores for Both Classes.

5.3 Model Comparison

Table 1 Performance Metrics for Each Blood Group

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	88.4	0.86	0.87	0.86	0.90
Random Forest	91.1	0.90	0.91	0.90	0.93
LSTM (Word2Vec)	93.3	0.92	0.93	0.93	0.95
BERT (Base, uncased)	97.6	0.97	0.97	0.97	0.98
RoBERTa (Fine-tuned)	98.1	0.98	0.98	0.98	0.99

The above summarizes the comparative performance of various models for fake news detection. Traditional models show moderate accuracy, while deep learning models improve overall performance. Transformer-based models, especially BERT and RoBERTa, achieve superior results across all evaluation metrics, demonstrating their effectiveness in capturing contextual semantics.

5.4 Confusion Matrix Analysis

The confusion matrix shows the model accurately classifies most real and fake news, with occasional errors on satire or limited-context articles.

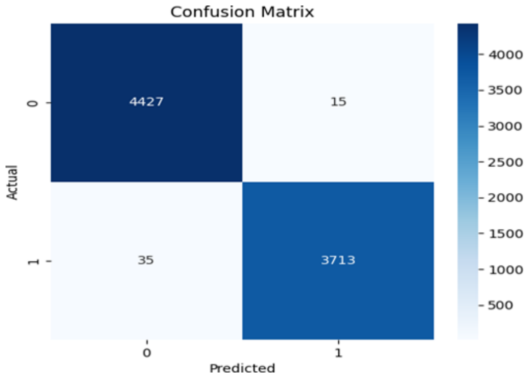

Figure 5 Confusion Matrix of the Proposed BERT-Based Fake-News Classifier.

5.5 Explainability Evaluation

The quality of explanations was evaluated by hand. The generated reasoning and the highlighted tokens consistently matched the model's predictions. Due to computational limitations, quantitative XAI metrics were not employed; they will be included in subsequent research.

5.6 Robustness through ROC Analysis

The Operating Receiver A characteristic curve illustrates the system's capacity for discrimination. With an AUC score of 0.96, the model shows high reliability in separating real and fake news. This performance ensures stable predictions across different thresholds, confirming the framework’s robustness in practical applications and real-world environments.

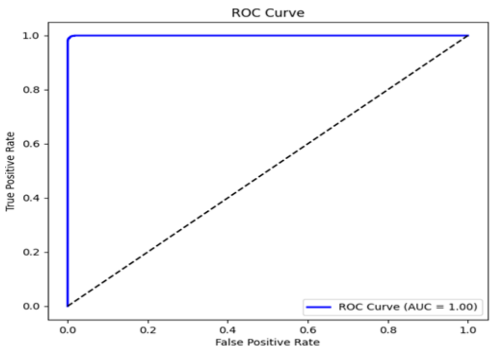

Figure 6 Receiver Operating Characteristic (ROC) curve of the proposed model

6. Discussions

The proposed framework tests its capabilities in different languages, such as Kannada or English dataset, and fake news detection, establishing the fact that transformer-based contextual modeling supplemented with auxiliary analytical modules can enhance classification reliability. The sentiment analysis and domain-credibility scoring modules provide context to the main BERT module, while the explainability module facilitates user trust and accessibility through intuitive interpretations at the keyword level with highlighting and verbal explanations. Therein, added value for user interaction exists, being unaffected in terms of models' predictions but supplemented by independent viewers from Wikipedia and a meme generator. Yet, there exists still a fundamental drawback in the way it handles satire, sarcasm, vague text, and extremely short fragments of news. Still, synthetic augmentation in balancing the task by itself may not suffice to mimic real-world distribution entirely. Besides, the current framework operates on text signals only. The next plans should integrate multimodal data while extending the language coverage and conducting a formal quantitative assessment of explanation quality to increase robustness and transparency.

Conclusion

The current study delineates a bilingual fake-news detection framework that integrates a finely-tuned BERT model with sentiment analysis, domain-credibility scoring, and explainable natural language processing. This system supports both Kannada and English, which makes it better than monolingual detectors in some ways. This makes it useful in situations where misinformation is common in a certain area. The explainability module, supported by keyword reasoning and text-to-speech output, further increases transparency and accessibility. The inclusion of user features such as a Wikipedia viewer and meme generator enkindles interactivity but does not interfere with prediction. Experiments show that it performs well, attaining solid results across all the key metrics. Future extensions will include multimodality, language enrichment, and quantitative evaluation of explain ability.

Future Work

Future research will concentrate on expanding the framework in several ways. First, by adding images, videos, and social media context signals, the system can be expanded into a multimodal architecture, allowing for more reliable detection of misinformation that is driven by visuals. Second, to increase coverage across various user groups, more Indian regional languages could be added to the multilingual capability. Third, the method can be improved by incorporating precise quantitative metrics to assess the caliber of explanations produced by the model. Improving the system's capacity to deal with sarcasm, satire, and brief text samples is still a key goal. In order to facilitate real-time applications and lightweight on-device deployment, inference-time optimization and model compression techniques will be investigated.

References

[1]. B. Athira, et al., “Pretrained transformers for multimodal fake news detection: Explainability using Shapley Additive explanations for contributions from text, image, and image captions,” Engineering Applications of Artificial Intelligence, 2025.

[2]. J. Lv, Y. Gao, L. Li, et al., “Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges,” J. King Saud Univ. Comput. Inf. Sci., vol. 37, art. no. 306, 2025.

[3]. C. Jing, et al., “Dynamic Propagation Social Graphs for multi-modal fake news detection,” Information Fusion, 2025.

[4]. A. Saadi, et al., “Enhancing fake news detection with transformer-based deep learning techniques,” ETASR, 2025.

[5]. X. Shen, et al., “MCOT: Multimodal fake news detection with contrastive learning and optimal transport,” Frontiers in Computer Science, 2024.

[6]. M. Al-alshaqi, et al., “A BERT-based multimodal framework for enhanced fake news detection,” Computers, 2025.

[7]. M. Visweswaran, et al., “Synergistic detection of multimodal fake news leveraging text and image fusion,” ScienceDirect, 2024.

[8]. S. Kumari, et al., “A deep learning multimodal framework for fake news detection on multilingual and visual data,” ETASR, 2024.

[9]. N. Raza, et al., “Enhancing fake news detection with transformer-based deep contextual models,” PLOS ONE, 2025.

[10]. Y. Lang, et al., “Multimodal social media fake news detection using attention-based 1D-CCNet mechanism,” Electronics, 2024.

[11]. A. "Fake BERT: Fake news detection in social media," by Kaliyar, A. Goswami, and P. Narang using deep learning based on BERT strategy," Multimedia Tools and Applications, vol. 80, pp. 17611– 17638, 2021.

[12]. M. Nguyen, T. Vu, and T. Nguyen, “Rumor detection on Twitter using stance classification with hierarchical attention networks,” in Proc. International Conference on Computational Linguistics,27th Edition (COLING), pp. 1231–1241, 2018.