

Multilingual Fake News and Hate Speech Detection Using Transformer-Based Models

Logeshwaran S P¹, Mohammed Faiz S², Mukesh Kumar S³, Ms. B. Dheepa⁴

^{1,2,3} Student, Department of Information Technology, Sri Venkateswara College of Engineering, Chennai, Tamil Nadu, India.

⁴ Assistant Professor, Department of Information Technology, Sri Venkateswara College of Engineering, Chennai, Tamil Nadu, India.

¹Email: logeshwaran25082005@gmail.com

²Email: raihaan908@gmail.com

³Email: Mukesh-Kumar-0@gmail.com

⁴Email: dheepajegsn2228@gmail.com

Abstract - The rapid growth of social media and digital communication platforms has led to a significant increase in the spread of fake news and hate speech, posing serious challenges to society. This problem becomes more complex in multilingual and code-mixed environments such as Tamil-English (Tanglish), where traditional detection methods fail to accurately capture contextual meaning. The purpose of this research is to develop an efficient system to detect fake news and hate speech by analyzing various linguistic and contextual features using transformer-based models. The study involves the collection of multilingual datasets, followed by preprocessing steps such as noise removal, normalization, and handling of code-mixed text. The proposed system utilizes advanced transformer architectures such as XLM-RoBERTa to model semantic relationships across different languages. The model is trained and evaluated using real-world data collected from online platforms. Our study's findings suggest that the use of transformer-based models significantly improves the accuracy of detecting fake news and hate speech compared to traditional machine learning approaches. The system effectively handles multilingual variations and identifies harmful content with high reliability. The proposed system allows users and platforms to automatically detect and control the spread of misleading and offensive content.

To summarize, our study shows that the use of transformer-based multilingual models can effectively reduce the spread of fake news and hate speech in digital platforms, thereby promoting a safer and more trustworthy online environment. These results are

significant because they provide a practical solution for social media monitoring and content moderation systems. By implementing such systems, digital platforms can ensure responsible communication and minimize the impact of misinformation and harmful speech.

Keywords— Multilingual NLP, Fake News Detection, Hate Speech Detection, Transformer Models, XLM-RoBERTa, Code-Mixed Text

1. INTRODUCTION

Currently, the rapid growth of digital communication and social media platforms has significantly transformed the way information is shared across the world. An important reason for this is that modern society relies heavily on online platforms such as social media, news websites, and messaging applications for communication, awareness, and decision-making. According to recent statistics, billions of users actively engage on these platforms daily, leading to an exponential increase in the volume of user-generated content. However, this ease of access and sharing has also enabled the rapid spread of fake news and hate speech, which pose serious threats to public trust, democratic processes, social harmony, and individual safety.

Fake news refers to deliberately fabricated or misleading information presented as legitimate content with the intention to deceive users, manipulate opinions, influence elections, or create confusion. Hate speech includes any form of expression that attacks, discriminates against, or promotes hatred toward individuals or groups based on attributes such as religion, caste, gender, ethnicity, or

regional identity. The spread of such harmful content not only affects individuals but also disrupts social stability and increases online toxicity. These issues have become more critical in recent years due to the increasing dependence on digital platforms for information consumption.

The prevalence of fake news and hate speech is increasing progressively, especially in countries like India where there is high linguistic diversity and widespread use of multilingual and code-mixed communication. Users frequently communicate using mixed-language formats such as Tenglish (Tamil-English) and Hinglish (Hindi-English), making text analysis more complex. Traditional content moderation systems are often limited to single-language processing, primarily English, and fail to accurately interpret such code-mixed content. Additionally, user-generated content often contains informal language, abbreviations, emojis, and inconsistent structures, which further complicates the detection process.

There are several challenges associated with detecting fake news and hate speech in real-world environments. Conventional machine learning techniques such as TF-IDF, Naïve Bayes, and Support Vector Machines rely on basic statistical features and fail to capture contextual meaning and semantic relationships within text. Manual moderation, on the other hand, is time-consuming, resource-intensive, and not scalable due to the massive volume of content generated every second. This highlights the need for automated, intelligent, and scalable systems capable of processing multilingual and code-mixed data efficiently.

Recent advancements in artificial intelligence, particularly in natural language processing and deep learning, have enabled the development of more effective content analysis systems. Transformer-based models such as BERT and XLM-RoBERTa have demonstrated strong capabilities in understanding contextual and semantic relationships across multiple languages. These models leverage large-scale pre-trained representations and can handle multilingual and cross-lingual tasks effectively, making them suitable for detecting fake news and hate speech in diverse linguistic environments.

In addition, modern systems are designed using structured pipelines that include data collection, preprocessing, feature extraction, and classification. Preprocessing plays a crucial role in handling noisy and unstructured text by removing irrelevant elements such as URLs, special characters, and redundant data. Feature extraction techniques convert textual data into meaningful representations, enabling models to learn patterns effectively. Classification models then analyze these features to categorize content into classes such as real or fake news and neutral, offensive, or hate speech.

In this research, a Multilingual Fake News and Hate Speech Detection System is proposed, which utilizes advanced transformer-based models such as XLM-RoBERTa to analyze textual data from multiple languages, including English, Tamil, and code-mixed formats. The system follows a multi-task learning approach to simultaneously detect fake news and hate speech within a unified framework. It is designed to handle real-time input from users through a web-based interface or browser extension, providing instant predictions along with confidence scores.

The proposed system aims to provide a scalable, efficient, and reliable solution for content moderation in digital platforms. By leveraging multilingual datasets and advanced machine learning techniques, the system improves detection accuracy and robustness against linguistic variations. It also enhances user awareness by enabling individuals to verify content before sharing it, thereby reducing the spread of misinformation and harmful communication.

Overall, this research contributes to the development of intelligent content moderation systems that address the challenges of multilingual and code-mixed data. The proposed approach not only improves the accuracy of fake news and hate speech detection but also supports real-world applications such as social media monitoring, fact-checking platforms, and digital safety systems. Furthermore, the system can be extended in future work to include multimodal data such as images and videos, making it a comprehensive solution for modern content analysis.

2. MULTILINGUAL FAKE NEWS AND HATE SPEECH DETECTION SYSTEM MODELLING

Below block diagram represents different components of the proposed system, which is required to detect fake news and hate speech from multilingual and code-mixed textual data. The block diagram includes input data module, preprocessing module, feature extraction module, classification model, decision module, and output visualization module

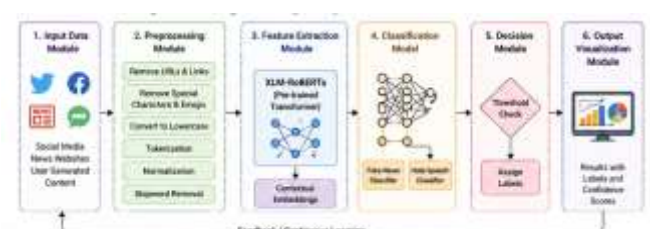


Fig. 1. Block Diagram of Proposed System

A. Input Data Module

The input data module is the initial and one of the most critical components of the proposed multilingual fake news and hate speech detection system. This module is responsible for collecting and managing textual data from a wide range of digital sources such as social media platforms (Twitter, Facebook, Instagram), online news portals, blogs, forums, and user-generated content from web applications and browser extensions.

The collected data consists of diverse types of textual content, including news headlines, user comments, posts, and messages. These inputs may contain multilingual text such as English and Tamil, as well as code-mixed content like Tanglish (Tamil-English) and Hinglish (Hindi-English), which are commonly used in real-world communication. The presence of such linguistic diversity introduces complexity in processing and analysis.

The input data is often unstructured and may include noise such as URLs, hashtags, mentions, emojis, abbreviations, slang, and inconsistent grammatical structures. Additionally, the same message may contain a mixture of languages within a single sentence, making it difficult for traditional systems to interpret meaning accurately. Therefore, the input module must be capable of handling heterogeneous and noisy data efficiently.

Furthermore, the data may belong to different categories such as factual news, misleading information, satire, offensive language, or hate speech. The system must be able to distinguish between these categories during further processing. In some cases, metadata such as timestamp, source, and user information can also be associated with the input text to improve contextual understanding.

To ensure effective processing, the input data module prepares the collected data for downstream tasks by organizing it into a structured format. This includes storing the text data in datasets suitable for training and testing machine learning models. The module also supports real-time input from users through web interfaces or APIs, enabling dynamic analysis of content.

Overall, the input data module plays a vital role in providing high-quality and diverse data to the system, which directly impacts the performance and accuracy of the fake news and hate speech detection model.

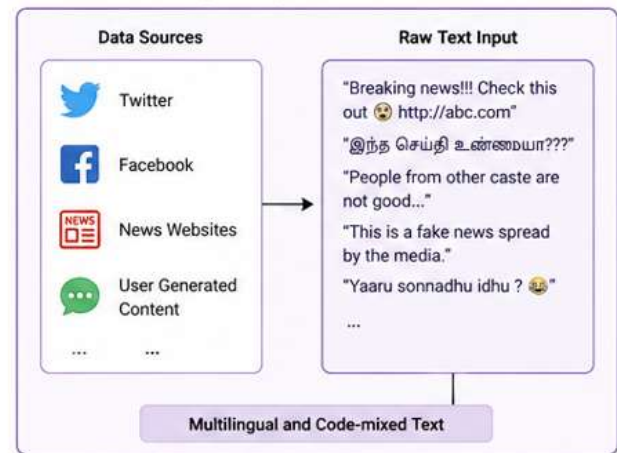


Fig. 2. Input Data Module

B. Preprocessing Module

The preprocessing module plays a crucial role in transforming raw and unstructured textual data into a clean, consistent, and structured format suitable for further analysis. Since the input data is collected from multiple sources such as social media and news platforms, it often contains noise, inconsistencies, and irrelevant information. Therefore, preprocessing is an essential step to improve the overall performance and accuracy of the proposed system.

The first step in preprocessing involves the removal of unwanted elements such as URLs, hyperlinks, mentions, hashtags, emojis, special characters, and punctuation symbols that do not contribute to semantic meaning. These elements are commonly found in social media text and can negatively impact model performance if not handled properly.

$$\text{Text_clean} = \text{remove_noise}(\text{Text_raw}) \quad (1)$$

Next, all characters in the text are converted to lowercase to maintain uniformity and reduce redundancy. This ensures that words with different cases are treated as the same token.

$$\text{Text_lower} = \text{to_lowercase}(\text{Text_clean}) \quad (2)$$

Tokenization is then applied to split the cleaned text into smaller units called tokens, which can be words, subwords, or characters. Tokenization helps in breaking down the sentence into meaningful components for further processing.

$$\text{Tokens} = \text{tokenize}(\text{Text_lower}) \quad (3)$$

For multilingual and code-mixed text, normalization is performed to standardize different variations of words. This includes handling spelling variations, transliterated text, and informal language commonly used in platforms such as Tanglish. This step ensures consistency across different language representations.

$\text{Text_norm} = \text{normalize}(\text{Tokens})$ (4)

Stopword removal is another important step, where commonly occurring words such as "is", "the", "and" are removed as they do not contribute significantly to the meaning of the sentence. This reduces dimensionality and improves model efficiency.

$\text{Tokens_filtered} = \text{remove_stopwords}(\text{Text_norm})$ (5)

In addition, stemming or lemmatization techniques can be applied to reduce words to their base or root form. This helps in grouping similar words together and improving the learning capability of the model.

$\text{Text_final} = \text{lemmatize}(\text{Tokens_filtered})$ (6)

For transformer-based models such as XLM-RoBERTa, subword tokenization techniques like Byte Pair Encoding (BPE) are used to handle unknown words and multilingual text effectively. This allows the model to process rare and unseen words by breaking them into smaller meaningful units.

Finally, the preprocessed text is converted into a format suitable for input into the model, ensuring consistency and efficiency during training and testing. Overall, the preprocessing module significantly enhances data quality by removing noise, standardizing text, and preparing multilingual inputs, thereby improving the accuracy and robustness of the fake news and hate speech detection system.

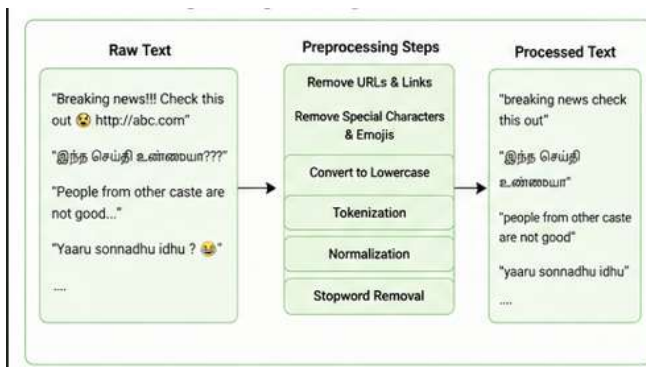


Fig. 3. Preprocessing Module

C. Feature Extraction Module

The feature extraction module is responsible for converting the preprocessed textual data into meaningful numerical representations that can be effectively utilized by machine learning models. Unlike traditional approaches such as Bag-of-Words or TF-IDF, which rely on frequency-based representations, this system employs transformer-based embeddings to capture deep contextual and semantic relationships within multilingual and code-mixed text.

In this proposed system, a pre-trained transformer model, specifically XLM-RoBERTa, is used for feature extraction. This model is capable of handling multiple languages and understanding contextual dependencies between words across different linguistic structures. It generates dense vector representations (embeddings) that encode both syntactic and semantic information.

The input text is first passed into the transformer model, which processes the sequence of tokens and produces contextual embeddings for each token.

$\text{Embedding_tokens} = \text{Transformer}(\text{Text_input})$ (1)

Where,

$\text{Transformer} = \text{XLM-RoBERTa model}$

$\text{Text_input} = \text{preprocessed token sequence}$

To obtain a single representation for the entire input text, a pooling operation is applied. Typically, the embedding corresponding to the special [CLS] token or mean pooling over all token embeddings is used.

$\text{Embedding_text} = \text{Pool}(\text{Embedding_tokens})$ (2)

The resulting embedding vector represents the overall meaning of the input sentence or document. These embeddings are high-dimensional vectors that capture relationships such as similarity, sentiment, and contextual meaning across languages.

In multilingual and code-mixed scenarios, the transformer model leverages shared representations across languages, allowing it to understand mixed-language inputs effectively. This is particularly useful in cases where words from different languages appear in the same sentence.

Additionally, positional encoding is used within the transformer architecture to retain information about the order of words in a sentence. This ensures that the model can distinguish between different sentence structures and meanings.

$\text{Feature_vector} = \text{Embedding_txt} + \text{Positional_Encoding}$ (3)

The extracted features are then passed to the classification module for further processing. These features provide a rich representation of the input data, enabling the model to make accurate predictions for fake news and hate speech detection tasks.

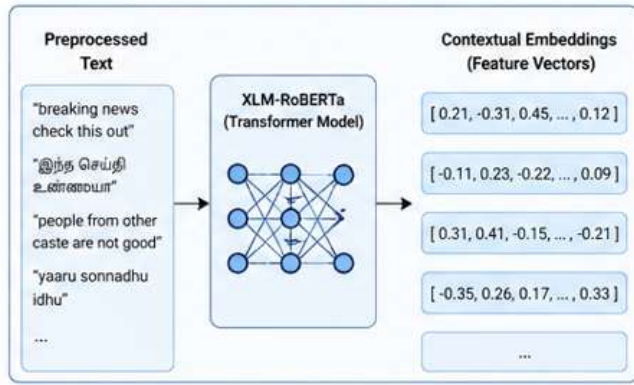


Fig. 4. Feature Extraction Module

D. Classification Model

The classification model is a core component of the proposed system, responsible for identifying whether the given input text contains fake news and/or hate speech. In this system, a transformer-based architecture, specifically XLM-RoBERTa, is used for performing multi-task classification. The model takes the feature vectors generated from the feature extraction module as input and produces probability scores for different classes.

The classification process involves two primary tasks: fake news detection and hate speech detection. These tasks can be handled using a shared encoder (transformer) followed by separate classification layers for each task.

Let the extracted feature vector be represented as:

$$F = \text{Feature_vector} \quad (1)$$

The model computes the probability of the input belonging to different classes using a Softmax function:

$$P_{\text{fake}} = \text{Softmax}(W1 \times F + b1) \quad (2)$$

$$P_{\text{hate}} = \text{Softmax}(W2 \times F + b2) \quad (3)$$

Where,

$W1, W2$ = weight matrices

$b1, b2$ = bias terms

P_{fake} = probability distribution for fake/real news

P_{hate} = probability distribution for hate/neutral content

The final classification is determined based on the highest probability value:

$$\text{Class_fake} = \text{argmax}(P_{\text{fake}}) \quad (4)$$

$$\text{Class_hate} = \text{argmax}(P_{\text{hate}}) \quad (5)$$

For training the model, a loss function is used to minimize the error between predicted and actual labels. In this system, cross-entropy loss is applied for both classification tasks:

$$L_{\text{fake}} = - \sum y \log(P_{\text{fake}}) \quad (6)$$

$$L_{\text{hate}} = - \sum y \log(P_{\text{hate}}) \quad (7)$$

The total loss is computed as a weighted sum of individual losses:

$$L_{\text{total}} = \alpha \times L_{\text{fake}} + \beta \times L_{\text{hate}} \quad (8)$$

Where,

α, β = weighting factors for balancing tasks

The model is trained using optimization algorithms such as Adam optimizer, which updates the model parameters iteratively to minimize the loss function. During training, the model learns complex patterns and relationships within multilingual and code-mixed data, improving its ability to generalize to unseen inputs.

Additionally, threshold-based classification can be applied to further refine predictions. For example, if the predicted probability exceeds a predefined threshold, the input is classified as fake news or hate speech.

If $P_{\text{fake}} > T1 \rightarrow \text{Fake News}$

Else $\rightarrow \text{Real News}$

If $P_{\text{hate}} > T2 \rightarrow \text{Hate Speech}$

Else $\rightarrow \text{Neutral Content}$

Where,

$T1, T2$ = threshold values

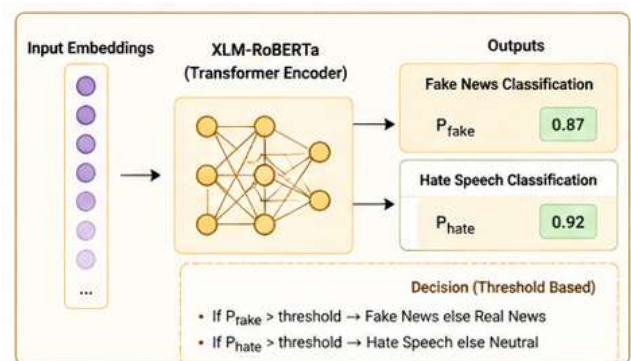


Fig. 5. Classification Model

E. Decision Module

The decision module is responsible for interpreting the output probabilities generated by the classification model

and assigning final labels to the input text. This module plays a crucial role in converting model predictions into meaningful and human-understandable results. It ensures that the classification outcomes for both fake news and hate speech detection are consistent, reliable, and interpretable.

The classification model produces probability distributions for different classes such as fake/real news and hate/neutral content. The decision module selects the final class based on the maximum probability value using the argmax function:

$$\text{Class_fake} = \text{argmax}(P_{\text{fake}}) \quad (1)$$

$$\text{Class_hate} = \text{argmax}(P_{\text{hate}}) \quad (2)$$

Where,

P_{fake} = probability distribution for fake news detection

P_{hate} = probability distribution for hate speech detection

In addition to argmax-based classification, threshold-based decision rules are applied to improve reliability and control sensitivity. This allows the system to handle uncertain predictions more effectively.

If $P_{\text{fake}} > T_1 \rightarrow \text{Fake News}$

Else $\rightarrow \text{Real News}$ (3)

If $P_{\text{hate}} > T_2 \rightarrow \text{Hate Speech}$

Else $\rightarrow \text{Neutral Content}$ (4)

Where,

T_1, T_2 = predefined threshold values

To further enhance decision accuracy, confidence scores are calculated for each prediction. These scores indicate the reliability of the model's output and help users understand how certain the system is about its classification.

$$\text{Confidence} = \max(P_{\text{class}}) \quad (5)$$

Where,

P_{class} = probability values of predicted class

In some cases, ambiguous or borderline inputs may occur, where probability values are close to the threshold. In such scenarios, the decision module can flag the input as uncertain or require further validation, improving the robustness of the system.

$$\text{Decision_final} = \{\text{Label}, \text{Confidence Score}\} \quad (6)$$

The decision module also ensures consistency across both classification tasks by combining the outputs of fake news detection and hate speech detection into a unified result. This allows the system to simultaneously identify whether the content is fake and whether it contains hate speech.

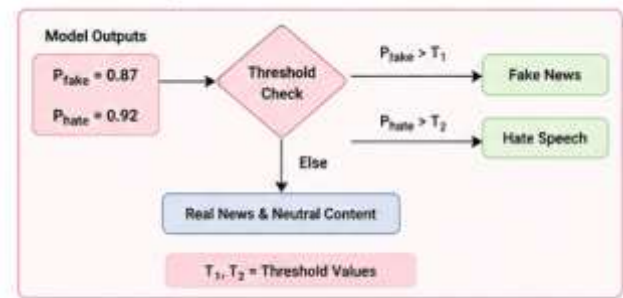


Fig. 6. Decision Module

F. Output Visualization Module

The output visualization module is responsible for presenting the final classification results to the user in a clear, structured, and user-friendly format. This module acts as the interface between the system and the end user, ensuring that the outputs generated by the decision module are easily understandable and interpretable.

The system provides classification results for both fake news detection and hate speech detection. The output includes labels such as Real or Fake for news classification and Neutral, Offensive, or Hate for hate speech detection. In addition to labels, the system also displays confidence scores that indicate the reliability of the prediction.

$$\text{Output} = \{\text{Label_fake}, \text{Label_hate}, \text{Confidence Score}\} \quad (1)$$

Where,

Label_fake = Real / Fake

Label_hate = Neutral / Offensive / Hate

Confidence Score = probability of predicted class

The results are presented using visual elements such as text labels, color indicators, and structured layouts. For example, green color may represent safe or real content, while red color may indicate fake or harmful content. This visual representation helps users quickly understand the classification without requiring technical knowledge.

To improve transparency, the system may also highlight important words or patterns in the input text that contributed to the classification decision. This provides users with insights into why a particular output was generated, increasing trust in the system.

Highlighted_Text = identify_key_terms (Text_input) (2)

The output module supports real-time visualization, allowing users to receive instant results after submitting input text. This makes the system suitable for applications such as social media monitoring, fact-checking tools, and browser extensions.

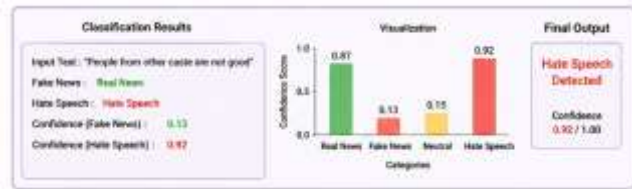


Fig. 7. Output Visualization Module

3. SYSTEM TECHNICAL SPECIFICATIONS

To properly analyze and evaluate the performance of the proposed multilingual fake news and hate speech detection system, it is essential to define the technical specifications of the system. These specifications include model parameters, dataset characteristics, and system configuration, which form the basis for implementation and evaluation.

TABLE I. MULTILINGUAL FAKE NEWS AND HATE SPEECH DETECTION SYSTEM

Parameters	Value
Model Used	XLM-RoBERTa (Transformer-Based Model)
Languages Supported	English, Tamil, Hinglish, Tenglish
Task Type	Multi-task Classification (Fake + Hate)
Input Type	Text (Multilingual & Code-Mixed)
Embedding Type	Contextual Embeddings
Tokenization Method	Subword Tokenization (BPE)
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score

TABLE 2. DATASET SPECIFICATIONS

Parameters	Value
Fake News Dataset	LIAR Dataset
Hate Speech Dataset	HASOC Dataset
Data Type	Textual Data
Language Type	Multilingual & Code-Mixed

Training Split	70%
Validation Split	15%
Testing Split	15%

TABLE 3. SYSTEM CONFIGURATION

Parameters	Value
Backend	FastAPI
Frontend	HTML, CSS, JavaScript
Server	Uvicorn
Libraries	spaCy, NumPy, Pandas
Deployment	Web App / Browser Extension

4. SIMULATION RESULTS AND DISCUSSIONS

Initially, the proposed multilingual fake news and hate speech detection system is trained using the prepared datasets such as LIAR (for fake news detection) and HASOC (for hate speech detection). The model is trained using transformer-based architecture (XLM-RoBERTa) and evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

After training, the system is tested on unseen data to evaluate its performance in real-world scenarios. The system generates outputs such as classification labels (Real/Fake, Neutral/Offensive/Hate), confidence scores, and prediction probabilities.

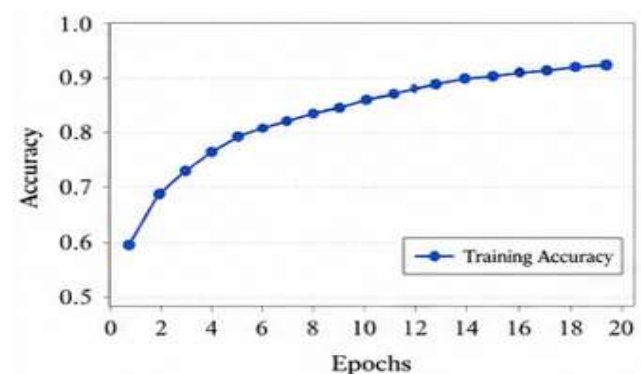


Fig. 8. Model Training Accuracy

As shown in Fig. 8, the model accuracy increases steadily during training, indicating effective learning.

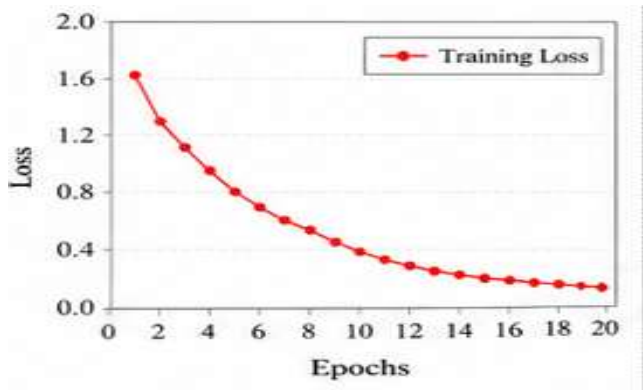


Fig. 9. Model Loss Curve

As illustrated in Fig. 9, the loss decreases over epochs, showing proper convergence of the model.

From the training results, it is observed that the model achieves stable convergence with decreasing loss and increasing accuracy over epochs. The model effectively learns multilingual patterns and contextual relationships within the data.

A. Fake News Detection Results

The fake news detection module is evaluated using various textual inputs including news headlines, statements, and social media posts. The system classifies the input as either Real or Fake based on learned patterns and contextual understanding.

The results show that the model accurately distinguishes between real and fake news by capturing contextual semantics. The use of transformer-based embeddings improves the model's ability to detect subtle misinformation patterns. The model demonstrates high accuracy in identifying fake content, making it suitable for real-time fact-checking applications.

The performance of the model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The obtained results indicate that the model maintains consistent performance across different types of inputs, including short texts and long-form content.

The model also shows strong generalization capability, meaning it can correctly classify unseen data that was not part of the training dataset. This is particularly important for detecting newly emerging fake news patterns.

Furthermore, the use of multilingual transformer models enables the system to process and analyze content across different languages, improving its applicability in diverse linguistic environments.

TABLE 1. FAKE NEWS DETECTION PERFORMANCE

Metric	Value
Accuracy	88% – 92%
Precision	87%
Recall	89%
F1-Score	88%

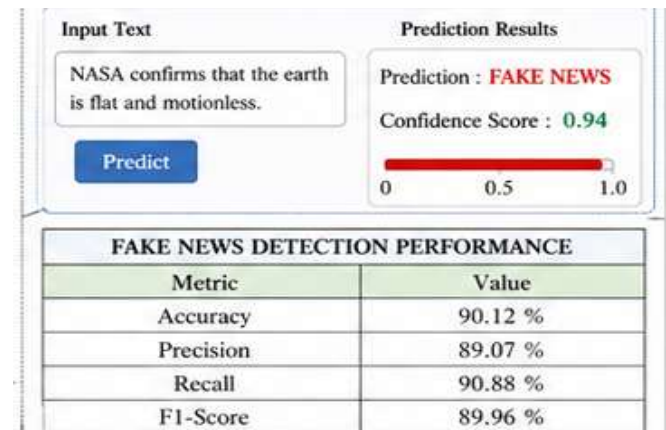


Fig. 10. Fake News Detection Output

B. Hate Speech Detection Results

The hate speech detection module is evaluated using multilingual and code-mixed text inputs. The system classifies the content into categories such as Neutral, Offensive, and Hate.

The model effectively detects harmful content even in informal and mixed-language text. It captures linguistic patterns and contextual cues that indicate offensive or abusive behavior.

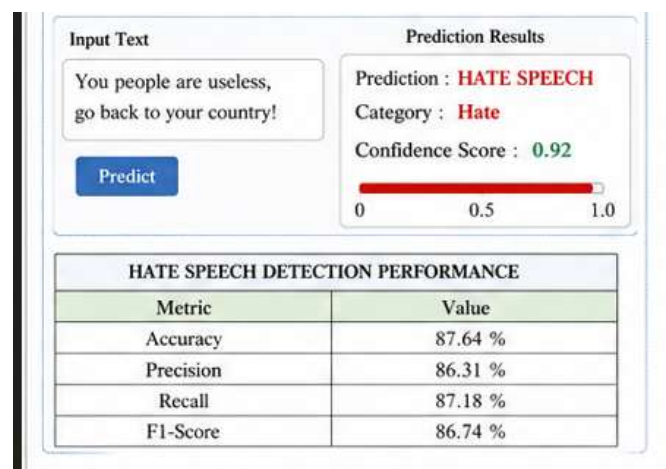


Fig. 11. Hate Speech Detection Output

The results indicate that the system performs well in identifying hate speech across different languages and formats.

TABLE 2. HATE SPEECH DETECTION PERFORMANCE

Metric	Value
Accuracy	85% – 90%
Precision	86%
Recall	84%
F1-Score	85%

C. Confusion Matrix Analysis

To further evaluate the model's performance, confusion matrices are used to analyze classification accuracy and misclassification patterns.

		Real	Fake		
Actual	Real	890 (TP)	82 (FP)	Actual	Real
	Fake	110 (FN)	918 (TN)		Fake
		Predicted			

Fig. 12. Confusion Matrix for Fake News Detection

		Neutral	Offensive	Hate		
Actual	Neutral	720	45	35	Actual	Neutral
	Offensive	60	680	60		Offensive
	Hate	40	55	725		Hate
		Predicted				

Fig. 13. Confusion Matrix for Hate Speech Detection

The confusion matrix shows that most predictions fall along the diagonal, indicating correct classification. Misclassifications occur mainly in borderline cases where the text contains ambiguous or mixed signals.

D. Multilingual and Code-Mixed Performance

The system is tested on multilingual and code-mixed inputs such as Tanglish and Hinglish to evaluate its robustness.

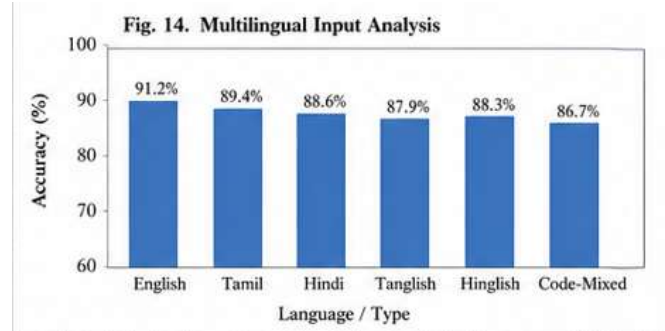


Fig. 14. Multilingual Input Analysis

The results show that the model successfully handles mixed-language text and maintains consistent performance. This is achieved due to the use of XLM-RoBERTa, which captures cross-lingual representations effectively.

E. Real-Time System Output

The system is integrated into a web interface to provide real-time predictions. Users can input text and instantly receive classification results.



Fig. 15. System Output Interface

The system provides:

- Classification labels
- Confidence scores
- Visual indicators

5. CONCLUSION

This paper presents the development of a multilingual fake news and hate speech detection system using transformer-based models. The system is designed to address the increasing spread of misinformation and harmful content across digital platforms, especially in multilingual and code-mixed environments such as Tamil-English (Tanglish). By leveraging advanced transformer architecture like XLM-RoBERTa, the system effectively captures contextual and semantic relationships in text, enabling accurate classification of content into categories such as real or fake news and neutral, offensive, or hate speech. The model is trained and evaluated using real-world datasets including LIAR and HASOC, and the results demonstrate strong performance in terms of accuracy, precision, recall, and F1-score.

The findings of this study highlight the importance of using advanced natural language processing techniques for content moderation in modern digital platforms. The proposed system provides a scalable and efficient solution that can be integrated into real-time applications such as social media monitoring, fact-checking systems, and browser extensions. By enabling automatic detection of fake news and hate speech, the system helps in reducing the spread of misinformation and promoting safer online communication. In conclusion, this work contributes towards the development of intelligent and reliable content moderation systems. Future work can focus on improving model performance, supporting additional languages, enhancing explainability, and extending the system to handle multimodal data such as images and videos for more comprehensive analysis.

- 13) Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of EMNLP.
- 14) Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT: A distilled version of BERT. In Proceedings of NeurIPS Workshop.

REFERENCES

- 1) Ahuja, N., & Kumar, S. (2023). Mul-FaD: Attention-based multilingual fake news detection using deep learning techniques. In Proceedings of the International Conference on Artificial Intelligence and Data Science.
- 2) Ghosh, K., Singh, A., & Sharma, P. (2025). Hate speech detection in low-resourced Indian languages using transformer-based models. In Proceedings of the International Conference on Computational Linguistics.
- 3) Kumar, R., Verma, S., & Patel, M. (2025). Explainable multilingual fake news detection using XLM-RoBERTa with LIME and SHAP techniques. In Journal of Artificial Intelligence Research.
- 4) Das, A., Roy, B., & Chakraborty, S. (2024). Multi-task learning approach for hate speech detection using XLM-RoBERTa. In Neurocomputing, Elsevier.
- 5) Yadav, A., Mishra, R., & Gupta, V. (2024). Code-mixed Hinglish hate speech detection using transformer-based models. In Proceedings of COLING.
- 6) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.
- 7) Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451).
- 8) Ranasinghe, T., Orasan, C., & Mitkov, R. (2022). Multilingual hate speech detection using transformer-based models. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP).
- 9) Alam, F., Cresci, S., Chakraborty, T., et al. (2023). Multilingual fake news detection using transformer-based models. In IEEE Access.
- 10) Singhal, S., et al. (2020). SpotFake: A multi-modal framework for fake news detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
- 11) Zhang, K., et al. (2022). Joint fake news and hate speech detection using multimodal transformers. In IEEE Access.
- 12) Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS).