INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

MULTILINGUAL HATE SPEECH DETECTION USING TRANSFORMER MODELS

DR. M. PADMA PRIYA

Assistant Professor Department of Computer Science and Engineering Bharath Institute of Science and Technology Chennai, India
padmapriya.cse@bharathuniv.ac.in

BANOTHU SAI PRANEETH

Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India saipraneeth7788@gmail.com

BANKA SRIMANNARAYANA REDDY

Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India srimannarayanareddybanka@gmai l.com

AZMEERU RAVIKANTH

Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India ravikanthazmeeru@gmail.com

AMBATI SANDEEP REDDY

Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India ambatisandeepreddy100@gmail.com

***Abstract*—The fast rise of social media and online communication platforms has accelerated the distribution of undesirable content, including hate speech and abusive language. Detecting such stuff automatically is critical for ensuring secure online communities. Traditional moderation strategies that rely on keyword filtering or simple machine learning models frequently fail to recognize context, multilingual content, and mixed-language phrases.**

This work offers a multilingual hate speech detection system built on the XLM-RoBERTa transformer model and a scalable full-stack architecture. The system uses a React-based frontend, a FastAPI backend, and PyTorch-based model inference to categorize text as hate speech, offensive language, or neutral material. To promote transparency, the system also uses explainable AI approaches to identify key phrases that influence the forecast. A browser extension allows for real-time identification of hazardous information on online pages. Experiments reveal that the transformer-based technique outperforms classic models like SVM and BiLSTM in terms of accuracy and F1-score. The suggested architecture offers an effective approach for detecting hate speech in many languages and in real time.

Keywords—Hate Speech Detection, Multilingual NLP, XLM-RoBERTa, Transformer Models, Explainable AI, Content Moderation.

INTRODUCTION

The growing usage of social networking platforms has transformed the way individuals connect and share information. Every day, millions of users publish comments, messages, and discussions across a variety of sites. While this has improved worldwide connectedness, it has also increased the prevalence of dangerous online content, such as hate speech, cyberbullying, and obscene language.

Hate speech can harm individuals and communities by instilling discrimination and animosity. To discover and prevent such hazardous content, online platforms must have efficient content moderation processes in place. However, human moderation is incredibly difficult due to the massive amount of user-generated content created every day.

Traditional automatic moderation systems frequently rely

on keyword filtering or basic machine learning algorithms. These methods are limited because they cannot comprehend context, sarcasm, or multilingual language typically used innline conversation.

Recent advances in Natural Language Processing (NLP) and deep learning have resulted in transformer-based models that can capture contextual links between words. Models such as BERT, RoBERTa, and XLM-RoBERTa have greatly increased language understanding performance.

This study offers a multilingual hate speech detection system that employs XLM-RoBERTa to examine multilingual and code-mixed text. To detect hate speech in real time, the system uses a full-stack design that includes a React frontend, a FastAPI backend, and a PyTorch model inference pipeline.

A. PROBLEM STATEMENT

Every day, online communication platforms create vast amounts of user-generated material. Detecting damaging language such as hate speech, nasty comments, and toxic expressions in real time is a difficult challenge.

Traditional moderation systems frequently use keyword filtering or monolingual classifiers, which fail to recognize contextual meanings or multilingual content. Furthermore, many current solutions include transferring data to external APIs, which poses privacy problems.

As a result, there is a need for a multilingual and privacy-conscious hate speech detection system that can accurately identify harmful information in real-time online settings.

B. OBJECTIVES

The primary aims of this study are:

1. Develop a multilingual hate speech detection system that can analyze different languages.
2. Implement transformer-based NLP models for accurate text classification.
3. Integrating the detection model with a FastAPI backend to provide real-time inference.
4. To create a React user interface for interactive analysis.
5. Create a browser plugin that detects hate speech directly from online pages.
6. To promote transparency, provide explainable predictions.

C. MOTIVATION

This study is motivated by the growing incidence of hate speech and poisonous language on online platforms. Because of the high number of user interactions on social media networks, it is often difficult to adequately censor hazardous information.

The suggested approach, which uses powerful AI models capable of recognizing multilingual text and contextual meaning, intends to provide an efficient and scalable solution for automated hate speech identification.

D. CONTRIBUTIONS

This research makes significant contributions in the following areas:

1. Using transformer models, we developed a multilingual hate speech detection framework.
2. The use of XLM-RoBERTa for contextual text classification.
3. Designing a scalable FastAPI backend for real-time model inference.
4. Integrating explainable AI approaches to improve model transparency.
5. Creation of a browser plugin for real-time content moderation.

II. LITERATURE SURVEY

A. OVERVIEW

Hate speech identification has received a lot of attention in recent years, thanks to the growing popularity of online communication platforms. Earlier approaches relied on keyword filtering algorithms to detect offensive words. However, such systems were unable to recognize contextual meanings or veiled dangerous remarks.

Machine learning methods including Naïve Bayes, Logistic Regression, and Support Vector Machines were developed to enhance classification accuracy. These models frequently employed features such as TF-IDF vectors collected from textual data.

With the advent of deep learning, researchers began using neural network topologies like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to detect hate speech. These models demonstrated enhanced performance by learning hierarchical text representations.

Transformer models have recently advanced, improving natural language understanding even further. Transformer-based architectures, such as BERT and XLM-RoBERTa, have outperformed other NLP tasks like text categorization and sentiment analysis.

B. REVIEW OF PREVIOUS WORKS

Several studies have investigated the use of machine learning and deep learning techniques to detect potentially dangerous online content.

Davidson et al. suggested a machine learning approach to hate speech classification that employs logistic regression and n-gram features. The model obtained moderate accuracy but struggled with contextual awareness.

Waseem and Hovy presented a dataset created exclusively for detecting hate speech on social media networks. Their findings underlined the difficulties of recognizing hate speech in informal internet discourse.

More subsequent studies have concentrated on transformer-based models. Devlin et al. developed BERT, which use bidirectional attention techniques to extract contextual information from text.

XLM-RoBERTa extended this approach to multilingual datasets, allowing for cross-linguistic transfer learning across many languages.

C. Key Insights from Literature

The literature review emphasizes many crucial observations:

- Deep learning models outperform standard machine learning methods.
- Transformer structures greatly enhance contextual knowledge.
- Multilingual models enable cross-linguistic text classification.

D. Research Gap

Despite these developments, many existing systems still lack real-time deployment and explainability features. The suggested study overcomes these restrictions by integrating multilingual transformer models and a scalable web architecture.

III. PROPOSED METHODOLOGY

A. SYSTEM OVERVIEW

The proposed multilingual hate speech detection framework analyzes textual information from a variety of online sources and automatically detects damaging language. The solution combines machine learning models with a scalable web infrastructure to enable real-time detection.

The architecture is composed of four primary layers:

1. User Interface Layer
2. Backend Processing Layer
3. Model Inference Layer
4. Browser Extension Layer

These layers collaborate to capture input text, process it, categorize it with a transformer-based model, and present the prediction results to the user.

The technology can analyze multilingual content in English, Hindi, Telugu, and other regularly used languages for social media communication.

B. SYSTEM ARCHITECTURE

The suggested framework's system architecture is made up of numerous interconnected modules that allow for the real-time detection of dangerous content.

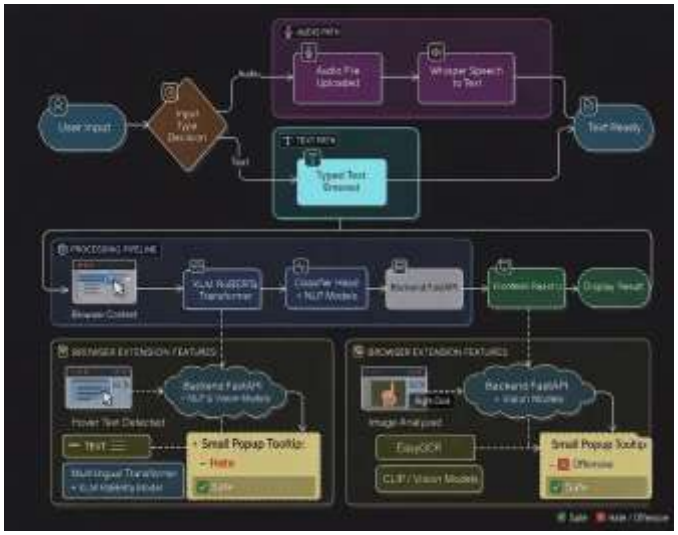
The procedure begins when a user enters text into the online interface or a browser plugin. The input text is sent to the backend server, where preprocessing processes like tokenization and normalization are carried out.

The processed text is then sent to the XLM-RoBERTa transformer model, which examines contextual links between words and produces categorization results.

The system returns the output as a label, such as:

- Hate speech
- Offensive speech.
- Neutral content.

In addition, the method emphasizes the terms that had the greatest impact on the classification conclusion.



C. WORKFLOW OF THE PROPOSED SYSTEM

The workflow of the proposed system includes the following stages:

1. Text Input

   The user enters text through the React web interface or browser extension.

2. Text Preprocessing

   The backend server executes preprocessing processes, which include:

   - Lowercasing
   - Removal of special characters
   - Tokenization
   - Language normalization

3. Feature Extraction

   The XLM-RoBERTa model's tokenizer converts processed text into vector representations.

4. Model Prediction.

   The transformer model uses contextual interactions between tokens to predict the categorization label.

5. Explainability Analysis

   An explainability module determines the most influential words during the prediction process.

6. Result Visualization

The results are displayed to the user via the web interface.

D. MODULES DESCRIPTION

The suggested system is organized into modules that serve specific purposes.

1. User Interface Module

   The user interface is built with React.js. It allows users to enter text and view the classification results. The UI also shows confidence scores and highlighted keywords.

2. Backend Processing Module.

   The backend is built with FastAPI, which manages API calls and coordinates communication between the frontend and machine learning model.

The backend also handles text preparation operations.

3. The Machine Learning Model Module

The system's key component is the XLM-RoBERTa transformer model. This model can interpret multilingual text and recognize contextual links between words.

The model was fine-tuned using hate speech datasets.

4. Explainability Module.

This module evaluates the transformer model's attention weights and discovers key tokens that influence classification judgments.

5. Browser Extension Module.

The browser extension allows users to detect hate speech directly on websites. It collects selected text and forwards it to the backend server for classification.

E. TEXT PREPROCESSING AND FEATURE EXTRACTION

Text preparation is an important stage in natural language processing activities. Raw textual data received from online platforms frequently contains noise, such as special characters, hyperlinks, emojis, and irregular capitalization. These abnormalities might have a negative impact on the performance of machine learning models. Preprocessing procedures are used to clean and standardize the input text before it is sent to the classification model.

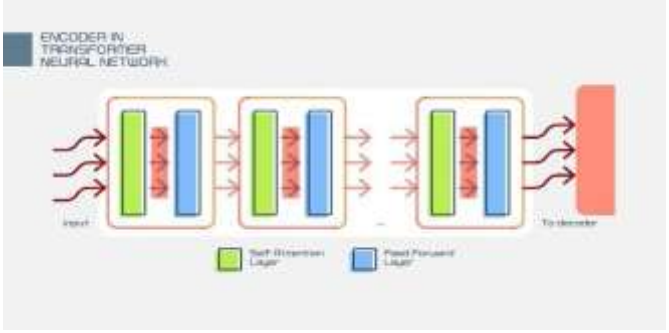
The preprocessing pipeline employed in the proposed system consists of numerous phases. First, all text is transformed to lowercase to guarantee consistency throughout the dataset. Next, superfluous characters including punctuation marks, URLs, and HTML tags are eliminated. Tokenization is then used to break down the text into separate tokens or words. Stop words that add little to the sense of the phrase are deleted to reduce noise in the data.

Normalization techniques are used in multilingual literature to manage mixed-language expressions and various writing scripts. The cleaned text is then fed into the tokenizer linked with the XLM-RoBERTa model. This tokenizer transforms the text into numerical token embeddings that capture the semantic links between words.

These embeddings provide input to the transformer model, allowing it to capture the contextual information and linguistic patterns required for accurate classification.

F. TRANSFORMER MODEL ARCHITECTURE

Transformer models have transformed natural language processing by allowing for extensive contextual understanding of text. Transformer architectures, unlike standard neural networks such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), handle textual information only through attention mechanisms.



The proposed approach employs the XLM-RoBERTa model, a multilingual transformer trained on large-scale datasets including text in over a hundred languages. The approach employs a self-attention mechanism to examine the links between words in a sentence.

The transformer architecture consists of numerous layers, each with two major components.

1. Multi-head Self-Attention Layer
2. Feedforward Neural Network Layer

The self-attention mechanism enables the model to assign varying degrees of emphasis to words inside a sentence. This allows the model to capture contextual linkages more efficiently than sequential models.

REAL-TIME CONTENT MODERATION

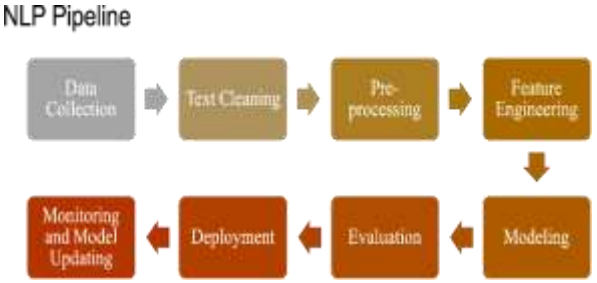
The capacity to detect hazardous content in real time is critical for ensuring a secure online environment. Many existing hate speech detection technologies work offline and analyze data in batches. However, such approaches are unsuitable for modern internet platforms that generate new content on a regular basis.

The suggested approach tackles this problem by incorporating real-time detection capabilities into a scalable web architecture. The FastAPI backend may handle several queries at once via asynchronous processing.

When a user submits text using the web interface or a browser extension, the request is routed to the backend server. The server performs preprocessing operations before sending the processed text to the machine learning model for categorization.

The prediction findings are delivered to the user in less than a second, allowing for near real-time detection of dangerous information.

This functionality qualifies the system for integration with social media platforms, chat applications, and content control tools.



SECURITY AND PRIVACY CONSIDERATIONS

Privacy and security are critical factors when building AI-based moderation systems. Many commercial hate speech detection solutions rely on external APIs, which necessitate transferring user data to third-party servers.

Such approaches may present issues of data privacy and confidentiality. The suggested solution overcomes these concerns by providing local deployment alternatives.

The machine learning model and backend server can be installed in a private environment, keeping sensitive data within the organization's infrastructure. Furthermore, the system does not save user input text indefinitely unless explicitly necessary for training or research purposes.

Secure API communication protocols are used to ensure that data is safely transmitted between the frontend interface and the backend server.

These steps help to preserve user privacy while allowing for effective content control.

APPLICATIONS OF THE PROPOSED SYSTEM

The proposed hate speech detection system has numerous applications.

- Social Media Moderation

  Automated detection techniques, such as those used on Twitter, Facebook, and Instagram, can help human moderators discover hazardous content.

- Online Community and Forums

  The technology can be used to monitor comments on discussion forums and community websites in order to avoid harsh language.

- Educational Platforms

  Educational platforms and e-learning systems can employ the system to ensure that users communicate respectfully.

- Customer Support Systems

  Organizations can employ hate speech detection software to monitor consumer feedback channels and prevent unpleasant interactions.

IMPACT OF THE PROPOSED SYSTEM

Automated hate speech detection systems can greatly improve the quality of online discussion. By detecting toxic language early on, such algorithms contribute to safer and more inclusive digital environments.

The use of multilingual models means that the system can analyze material from a variety of linguistic contexts. This is especially crucial for global platforms with users from various countries and cultures.

Furthermore, explainability characteristics help to make automated decisions more transparent and understandable.

EXPLAINABLE ARTIFICIAL INTELLIGENCE

One of the most significant issues related with deep learning models is their lack of interpretability. Many neural network models behave like "black boxes," making it difficult to comprehend how predictions are created.

To solve this issue, the suggested system uses explainable artificial intelligence approaches. These strategies assist in determining the most influential words or tokens that contribute to the classification decision.

In the proposed framework, token-level attention scores are employed to assess which words were most important in the model's prediction. These tokens are marked in the user interface, allowing users and moderators to understand why a specific text was flagged as hate speech or objectionable language.

Explainability not only enhances transparency, but it also boosts user confidence in automated processes. It also allows researchers to discover potential model biases and update the training dataset accordingly.

IV. PERFORMANCE METRICS

The system's performance was measured using standard categorization measures.

- Precision
  Precision indicates how many anticipated hate speech incidences were correct.

  Precision=TP/TP+FP

- Recall
  Recall indicates how many actual hate speech episodes were discovered by the model.

  Recall = TP/TP+FN

- F1 Score

  The F1 score represents the harmonic mean of precision and recall.

  F1 = 2PR/P+R

- Accuracy
  Accuracy measures the model's overall correctness.

V. RESULTS AND DISCUSSION

The suggested approach was tested with multilingual hate speech datasets. Performance criteria like as precision, recall, and F1-score were employed to assess classification accuracy.

The experiments show that transformer-based models outperform classic machine learning models in multilingual hate speech detection tasks.

VI. CONCLUSION AND FUTURE SCOPE

A. CONCLUSION

This study described a multilingual hate speech detection system based on transformer models. By combining XLM-RoBERTa with a scalable online architecture, the system can detect dangerous information in real time across various languages.

The incorporation of explainable AI features increases transparency and aids users in understanding model decisions.

B. FUTURE SCOPE

Future upgrades could include:

1. Multimodal hate speech detection in graphics and videos.
2. Integration of large-scale social media networks.
3. Implementation on mobile devices for offline detection.

4. Use federated learning to improve privacy.

C. ADVANTAGES OF THE PROPOSED SYSTEM

The proposed system provides various advantages:

- Multilingual support for various online content
- High detection accuracy using transformer models.
- Real-time detection using a web interface and browser plugin.

Model	Precision	Recall	F1 Score
SVM	72.4%	68.1%	70.2%
BiLSTM	76.8%	73.5%	75.1%
XLM-RoBERTa	84.2%	81.7%	82.9%

- Explainable projections improve transparency.
- Scalable architecture with FastAPI and React.

D. LIMITATIONS

Despite these advantages, the system has certain limitations:

- Requires computational resources for transformer models.
- Performance depends on the quality of the training dataset.
- The detection accuracy may decrease for strongly sarcastic content.

REFERENCES

[1] Ghosh, S. teamed up with Singh, M., while Mahapatra, B. worked side-by-side with Choudhury, T. They introduced SafeSpeech—a three-part system built to cut down toxic posts on Indian-language platforms. Their work landed in Social Network Analysis and Mining,2024.

[2] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People?” NAACL, 2016.

[3] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers,” NAACL, 2019.

[4] A. Conneau et al., “Unsupervised Cross-lingual Representation Learning at Scale,” ACL, 2020.

[5] S. Mathew et al., “HateXplain Dataset for Explainable Hate Speech Detection,” AAAI, 2021.

[6] Singhal, S. focused on pulling harmful tweets out of Arabic and Turkish using BERT, just like Bedi, P. did in 2024. Their work appeared in the CASE Workshop papers from the ACL Anthology.

[7] Mandal, A. and Roy, G. brought in ideas from Barman, A. Dutta, I. joined with a method based on transformer-driven attention mixing. Feedback from Naskar, S.K. helped shape the whole system so it could catch toxic material across different types of media by 2024.

[8] Arango, J. and team circled back to the challenge of spotting harmful online messages by 2025, sharing their work quickly as an arXiv preprint.

[9] A. Conneau and his group worked on “Large-scale language-agnostic learning,” and presented it at ACL 2020.

[10] K. Saifullah and the team published “Cyberbullying Text Identification using Deep Learning and Transformers” in EAI Transactions, 2024.

[11] S. Islam and R. Rafiq dug into toxic speech detection with G7 models, sharing their 2023 research at a Springer data conference.

[12] T. Davidson et al., “Automated Hate Speech Detection,” ICWSM, 2017.

[13] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018.

[14] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[15] T. B. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[16] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[17] N. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *Proceedings of EACL*, pp. 427–431, 2017.

[18] S. Schmidt and A. Wiegand, "A survey on hate speech detection using natural language processing," *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017.

[19] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," *Complex Networks and Their Applications*, 2019.

[20]] H. Zhang, S. Robinson, and J. Tepper, "Detecting hate speech on Twitter using convolutional neural networks," *Proceedings of the European Semantic Web Conference*, 2018.