iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81370

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Guardify: Intelligent Bilingual Cyberbullying Detection in Code-Mixed Hinglish Using MuRIL

Hashir Ahmad, Harsh Singh, Ibrahim Khan, Ms. Ankit Singh
Information Technology Babu Banarasi Das Institute of Technology and Management Lucknow, India

Abstract: ***The exponential proliferation of user- generated content on social media platforms has precipitated a parallel rise in cyberbullying, a digital menace with profound psychological and societal implications. While automated detection systems leveraging Natural Language Processing(NLP)haveachievedmaturityfor monolingual English corpora, they exhibit significantdegradationinperformancewhen applied to multilingual and code-mixed environments. In the Indian subcontinent, online communication predominantly features "Hinglish," a linguistic hybrid blending Hindi syntax with English vocabulary, characterized by non-standard orthography, fluid code-switching, and transliteration. This paper presents "Guardify," a comprehensive research framework and automated detection system specificallyengineeredforbilingual(English and Hinglish) cyberbullying detection.***

Moving beyond traditional lexicon-based approaches and shallow machine learning architectures, this study proposes the utilization of MuRIL (Multilingual Representations for Indian Languages), a BERT-based transformer model pre-trained on transliterated Indian data. We rigorously evaluate the proposed architecture against classicalbaselines(SVM,LogisticRegression) and sequential deep learning models (Bi- LSTM, CNN) using benchmark datasets includingtheBohraetal.corpusandHASOC shared tasks. The analysis demonstrates that the MuRIL-based approach significantly outperforms monolingual baselines in handling the semantic ambiguities of code- mixed text, offering a robust solution for content moderation in linguistically diverse digital ecosystem.

General Terms: Natural Language Processing, Machine Learning, Deep Learning, Text Classification, Cybersecurity, Social Media Analysis.

Keywords: Cyberbullying Detection, Hinglish, Code-Mixed Text, MuRIL, Transformer Models, Bilingual NLP, Hate Speech Detection, Text Classification, Multilingual Processing, Social Media Moderation.

I. INTRODUCTION

The advent of Web 2.0 has transformed the way information is created and shared, enabling unprecedented global connectivity and interaction. However, this digital expansion has also introduced serious challenges, particularly the rapid spread of offensive content, hate speech, and cyberbullying. Unlike traditional forms of bullying, which are limited by physical presence and time, cyberbullying is continuous, widespread, and often anonymous in nature. Its psychological impact on victims can be severe, leading to anxiety, depression, academic decline, and in extreme cases, long-term trauma or self-harm. As a result, the detection and prevention of such harmful content has become a critical concern for social media platforms, educational institutions, and policymakers striving to maintain safe digital environments.

The massive volume of user-generated content produced daily—ranging from tweets and comments to posts and messages—makes manual moderation impractical. This has driven the need for automated systems capable of identifying abusive and harmful content with high accuracy. Early approaches relied on keyword filtering and basic statistical techniques, but these methods lacked contextual understanding and were easily bypassed. With advancements in machine learning and deep learning, modern systems are now able to analyze context, semantics, and sentiment more effectively, enabling more reliable detection of cyberbullying.

A significant challenge in this domain arises from the multilingual nature of online communication, particularly in countries like India, where users frequently employ code-mixed languages such as Hinglish—a combination of Hindi and English. In such scenarios, users often switch between languages within the same sentence, with Hindi words written in Roman script alongside English text. This phenomenon introduces linguistic complexity, including inconsistent spelling, transliteration variations, and mixed grammatical structures. For instance, a phrase like “Tumhara attitude bahut burahaiyaar” combines Hindi and English seamlessly, making it difficult for conventional Natural Language Processing (NLP) systems to interpret.

©IJRASET: All Rights are Reserved

Traditional NLP models, primarily trained on standard English datasets, struggle to handle such code-mixed inputs. The absence of standardized spelling leads to multiple variations of the same word, significantly increasing vocabulary size and reducing model efficiency. Additionally, lexical ambiguity arises when words share similar forms across languages but carry different meanings. Furthermore, English-trained embeddings often fail to recognize transliterated Hindi words, treating them as unknown tokens and thereby losing critical semantic information required for accurate classification.

To address these challenges, this research proposes Guardify, an intelligent cyberbullying detection system specifically designed for bilingual English–Hinglish content. The system integrates a robust preprocessing pipeline tailored for code-mixed text with advanced machine learning and transformer-based models. By incorporating multilingual representations such as MuRIL, the proposed approach aims to capture the semantic nuances of Hinglish more effectively than traditional methods. This work presents a comprehensive comparison of classical machine learning models, deep learning architectures, and transformer-based approaches, evaluates performance on benchmark datasets, and explores the feasibility of deploying the system in real-world social media moderation environments.

II. LITERATURE REVIEW

Research on cyberbullying detection has evolved significantly over the past decade, progressing from simple rule-based systems to advanced deep learning and transformer-based approaches. Early studies primarily relied on lexicon-based filtering and classical machine learning techniques such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression. These models used features like bag-of-words and n-grams to identify abusive content. While computationally efficient and easy to interpret, such approaches were highly dependent on predefined vocabularies and struggled with variations in language, especially in informal and noisy social media text.

With the introduction of word embeddings such as Word2Vec and GloVe, researchers began incorporating semantic relationships between words into cyberbullying detection models. This advancement enabled better generalization compared to traditional feature-based methods. Subsequently, deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM), were applied to capture contextual and sequential information in text. These models demonstrated improved performance by learning patterns beyond surface-level keywords, especially in detecting implicit or context-dependent abusive language.

However, most of these models were designed for monolingual English datasets and showed limited effectiveness in multilingual or code-mixed scenarios. In the context of Indian social media, where users frequently communicate in Hinglish (a mixture of Hindi and English written in Roman script), traditional NLP systems face significant challenges. Code-mixed text introduces issues such as inconsistent spelling, transliteration variations, and frequent language switching within a single sentence. As a result, models trained on standard English corpora often fail to interpret the semantic meaning of such inputs accurately.

To address these limitations, researchers began exploring multilingual and cross-lingual approaches. Multilingual BERT (mBERT) extended transformer-based architectures to support multiple languages, providing contextual embeddings across different linguistic settings. While mBERT improved performance in multilingual tasks, it was still not specifically optimized for code-mixed Indian languages. This gap led to the development of specialized models such as MuRIL (Multilingual Representations for Indian Languages), which is pre-trained on Indian language corpora, including transliterated text. MuRIL has demonstrated superior performance in handling Hinglish and other code-mixed languages due to its ability to capture linguistic nuances and transliteration patterns.

Recent studies highlight that transformer-based models significantly outperform both classical and deep learning approaches in cyberbullying detection tasks. In particular, models fine-tuned on domain-specific datasets achieve higher accuracy, recall, and F1-scores. Additionally, preprocessing techniques such as transliteration normalization, emoji handling, and noise removal have been shown to further enhance model performance in real-world applications.

Despite these advancements, challenges remain in accurately detecting sarcasm, context-dependent abuse, and evolving slang in social media content. This research builds upon existing work by combining effective preprocessing strategies with a transformer-based approach using MuRIL, aiming to improve detection performance specifically for bilingual English–Hinglish text. The models were evaluated on a combined dataset derived from HASOC and code-mixed Hinglish corpora.

III. METHODOLOGY

The proposed system, Guardify, is designed as an end-to-end pipeline for detecting cyberbullying in bilingual English–Hinglish text.

The system integrates data preprocessing, feature extraction, and classification using both traditional machine learning and advanced transformer-based models. The methodology is structured into five major stages: dataset preparation, preprocessing, feature representation, model development, and evaluation.

The first stage involves dataset preparation. A composite dataset is created by combining publicly available datasets such as HASOC (Hate Speech and Offensive Content), TRAC (Trolling, Aggression, and Cyberbullying), and a custom Hinglish dataset collected from social media platforms such as Twitter, YouTube comments, and online forums. The dataset consists of labeled instances categorized into bullying and non-bullying classes. To ensure ethical compliance, all personally identifiable information is removed. The dataset is balanced to avoid bias, as real-world data often contains more non-abusive samples than abusive ones.

The second stage is preprocessing, which plays a critical role due to the noisy and informal nature of social media text. The preprocessing pipeline consists of multiple steps. First, noise removal is performed by eliminating URLs, user mentions, hashtags, and irrelevant symbols. Next, normalization is applied to reduce repeated characters (e.g., “stuuuupid” to “stupid”) and standardize text. Emoji handling is performed by converting emojis into textual representations (e.g., □ to “angry”) to preserve sentiment information. Since Hinglish involves transliteration, a normalization step is introduced to handle multiple spellings of the same Hindi word written in Roman script. Language identification is optionally applied at the token level to distinguish between English and Hindi words. Tokenization is then performed using whitespace tokenization for classical models and subword tokenization (WordPiece) for transformer models such as MuRIL.

The third stage focuses on feature representation. For classical machine learning models, Term Frequency–Inverse Document Frequency (TF-IDF) is used to convert text into numerical feature vectors based on word importance. For deep learning models, word embeddings such as Word2Vec or GloVe are used to capture semantic relationships between words. For transformer-based models, contextual embeddings are generated directly using MuRIL, which is pre-trained on Indian languages and code-mixed text. The contextual embeddings capture both semantic meaning and contextual dependencies within the sentence.

The fourth stage involves model development and training. Multiple models are implemented to compare performance across different approaches. Classical models such as Support Vector Machine (SVM) and Logistic Regression serve as baselines due to their simplicity and efficiency. A Bidirectional Long Short-Term Memory (Bi-LSTM) model is used to capture sequential dependencies and contextual relationships in text. The proposed model uses a fine-tuned MuRIL transformer, where the input text is passed through multiple self-attention layers to generate a contextual representation. The output corresponding to the classification token is fed into a fully connected dense layer followed by a softmax activation function to classify the text as bullying or non-bullying.

The fifth stage is training and optimization. The dataset is split into training, validation, and testing sets using a stratified approach to maintain class distribution. The models are trained using optimizers such as Adam or AdamW with appropriate learning rates. Hyperparameters such as batch size, sequence length, and dropout rate are tuned for optimal performance. Early stopping is implemented to prevent overfitting and ensure generalization.

Finally, the evaluation stage measures the performance of the models using standard metrics such as accuracy, precision, recall, and F1-score. Since cyberbullying detection is a sensitive task, recall and F1-score are prioritized to minimize false negatives. Cross-validation is performed to ensure robustness and reliability of the results. The proposed methodology effectively combines preprocessing techniques with advanced transformer-based learning to address the challenges of code-mixed bilingual text.

IV. SYSTEMARCHITECTURE

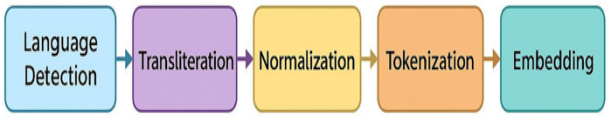


Figure 1: High-level system architecture for Data preprocessing pipeline.

The proposed system, Guardify, is designed as a modular and scalable architecture for detecting cyberbullying in bilingual English–Hinglish social media text. The architecture follows a pipeline-based approach, where each component performs a specific function, enabling efficient processing of noisy and code-mixed input data. The system consists of five major layers: Input Layer, Preprocessing Layer, Feature Extraction Layer, Model Layer, and Output & Evaluation Layer.

The Input Layer is responsible for collecting raw textual data from various social media sources such as Twitter, YouTube comments, and online discussion platforms. The input data consists of informal, unstructured, and often noisy text that may include slang, abbreviations, emojis, and code-mixed language patterns. This layer acts as the entry point of the system and ensures that the data is passed in a consistent format for further processing.

The Preprocessing Layer performs essential cleaning and normalization operations to improve the quality of the input text. This includes removal of URLs, user mentions, hashtags, and irrelevant special characters. Text normalization is applied to reduce repeated characters and correct inconsistent spellings. Since Hinglish text involves transliteration, this layer also handles multiple spelling variations of the same Hindi word written in Roman script. Additionally, emojis are converted into textual representations to preserve sentiment information. Tokenization is then performed to break the text into smaller units suitable for further analysis. For transformer-based models, subword tokenization techniques are used to handle unknown and rare words effectively.

The Feature Extraction Layer converts the processed text into numerical representations that can be used by machine learning models. For classical models, Term Frequency–Inverse Document Frequency (TF-IDF) is used to represent the importance of words in the document. For deep learning approaches, word embeddings are utilized to capture semantic relationships between words. In the proposed system, the MuRIL transformer model is used to generate contextual embeddings that capture both semantic meaning and contextual dependencies, making it highly effective for code-mixed language processing.

The Model Layer is responsible for classification. Multiple models are implemented and evaluated to identify the most effective approach. Classical machine learning models such as Support Vector Machine (SVM) and Logistic Regression serve as baseline models. A Bidirectional Long Short-Term Memory (Bi-LSTM) model is used to capture sequential dependencies in text. The core of the proposed system is the MuRIL-based transformer model, which uses self-attention mechanisms to understand contextual relationships across the entire input sequence. The output representation is passed through a fully connected layer followed by a softmax function to classify the text into bullying or non-bullying categories.

The Output and Evaluation Layer produces the final classification result along with performance metrics. The output indicates whether the given input text is classified as bullying or non-bullying. The system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Since detecting harmful content is critical, emphasis is placed on achieving high recall and F1-score to minimize false negatives. This layer also enables performance comparison between different models used in the system.

Overall, the system architecture is designed to effectively handle the challenges of multilingual and code-mixed text by integrating robust preprocessing techniques with advanced transformer-based learning. The modular design ensures flexibility, scalability, and adaptability for real-world deployment in social media moderation systems.

V. FLOW CHART OF THE WORKFLOW

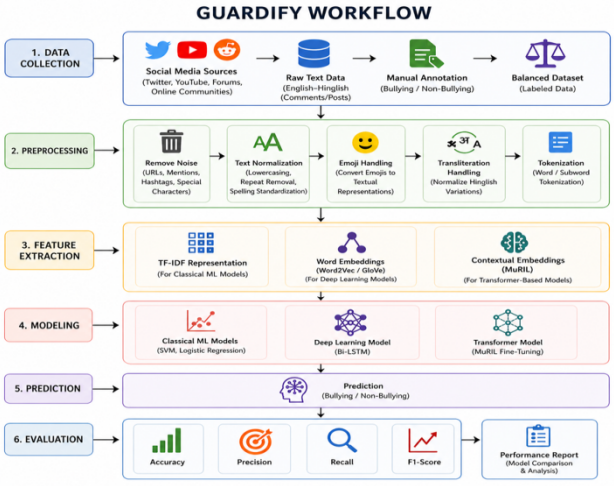


Figure 2: Flowchart of proposed model

A. Data Collection

The first stage of the workflow involves collecting raw textual data from various social media platforms such as Twitter, YouTube, Reddit, and online discussion forums. This data typically consists of user-generated content in the form of comments, posts, and messages, which may contain abusive or non-abusive language. The collected data is often bilingual in nature, combining English and Hinglish, and is highly unstructured. After collection, the data is manually annotated into predefined categories such as bullying and non-bullying to create a labeled dataset suitable for supervised learning. Ensuring diversity and balance in the dataset is important to improve model performance and reduce bias.

B. Preprocessing

The preprocessing stage focuses on cleaning and preparing the raw text for further analysis. Since social media data is noisy, this step removes unwanted elements such as URLs, user mentions, hashtags, and special characters. Text normalization is applied to standardize the data by converting text to lowercase, removing repeated characters, and correcting inconsistent spellings. Emojis are converted into textual representations to preserve emotional context. Additionally, transliteration handling is performed to normalize Hinglish words written in Roman script with multiple spelling variations. Finally, tokenization is carried out to break the text into smaller units (words or subwords), making it suitable for feature extraction.

C. Feature Extraction

In this stage, the processed text is transformed into numerical representations that machine learning models can understand. Different techniques are used depending on the model type. For classical machine learning approaches, TF-IDF (Term Frequency–Inverse Document Frequency) is used to represent the importance of words in a document. For deep learning models, word embeddings such as Word2Vec or GloVe are used to capture semantic relationships between words. For transformer-based models, contextual embeddings are generated using MuRIL, which captures both meaning and context of words in bilingual and code-mixed text.

D. Modeling

The modeling stage involves training different algorithms to classify the text as bullying or non-bullying. Classical machine learning models such as Support Vector Machine (SVM) and Logistic Regression are used as baseline approaches. A deep learning model, specifically Bidirectional Long Short-Term Memory (Bi-LSTM), is implemented to capture sequential patterns and contextual dependencies in text. The primary model used in this system is the MuRIL transformer, which leverages self-attention mechanisms to understand the full context of the sentence. These models are trained on the prepared dataset to learn patterns associated with cyberbullying.

E. Prediction

Once the models are trained, they are used to make predictions on unseen data. The input text is passed through the trained model, which outputs a classification label indicating whether the content is bullying or non-bullying. This stage is critical for real-world applications, as it determines how effectively the system can identify harmful content in real-time scenarios. The prediction output can also include confidence scores representing the probability of each class.

F. Evaluation

The final stage evaluates the performance of the models using standard metrics such as accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, while precision evaluates how many predicted bullying instances are actually correct. Recall is particularly important in this context, as it measures the ability of the system to detect all abusive content. F1-score provides a balance between precision and recall. The evaluation results are used to compare different models and identify the most effective approach for cyberbullying detection. A performance report is generated to summarize the findings and guide further improvements.

VI. IMPLEMENTATION DETAILS

The proposed system, Guardify, is implemented using Python due to its extensive support for machine learning and natural language processing libraries. The development environment includes Jupyter Notebook and Google Colab, enabling efficient experimentation and access to GPU acceleration for training deep learning and transformer-based models.

©IJRASET: All Rights are Reserved

For data handling and preprocessing, libraries such as NumPy and Pandas are used for data manipulation and cleaning. Text preprocessing is performed using standard Python libraries along with Natural Language Toolkit (NLTK) and regular expressions. This includes removal of URLs, mentions, and special characters, as well as text normalization and tokenization. Emoji handling is implemented using emoji libraries that convert symbols into textual descriptions. For handling code-mixed Hinglish text, custom normalization rules are applied to reduce spelling variations and improve consistency.

Feature extraction for classical models is performed using the Scikit-learn library, specifically utilizing the TF-IDF vectorizer to convert text into numerical feature representations. For deep learning models, word embeddings such as Word2Vec or GloVe are integrated to capture semantic relationships. The embeddings are either pre-trained or trained on the dataset, depending on availability and performance requirements.

The Bidirectional Long Short-Term Memory (Bi-LSTM) model is implemented using TensorFlow and Keras. The architecture consists of an embedding layer followed by a Bi-LSTM layer with a predefined number of hidden units, dropout layers for regularization, and a dense output layer with a sigmoid or softmax activation function for classification. The model is trained using the Adam optimizer and binary cross-entropy loss function.

For transformer-based modeling, the MuRIL model is implemented using the Hugging Face Transformers library. The pre-trained MuRIL model is fine-tuned on the prepared dataset for binary classification. The input text is tokenized using the MuRIL tokenizer with a maximum sequence length constraint. The tokenized inputs are passed through the transformer encoder, and the output corresponding to the classification token is used as the sentence representation. A fully connected dense layer is added on top, followed by a softmax activation function to produce classification probabilities.

The dataset is divided into training, validation, and testing sets using a stratified split to maintain class balance. Hyperparameters such as learning rate, batch size, and number of epochs are tuned to achieve optimal performance. Early stopping is implemented based on validation loss to prevent overfitting. Training is conducted using GPU acceleration where available to reduce computational time.

Evaluation is performed using Scikit-learn metrics, including accuracy, precision, recall, and F1-score. The performance of classical, deep learning, and transformer-based models is compared to identify the most effective approach. The results are visualized using Matplotlib and Seaborn to provide clear insights into model performance.

VII. EXPERIMENTAL RESULTS

The performance of the proposed Guardify system was evaluated using a combination of classical machine learning models, deep learning architectures, and transformer-based approaches on a bilingual English–Hinglish dataset. The dataset was divided into training, validation, and testing sets using a stratified split to maintain class balance. All models were trained and evaluated under the same conditions to ensure a fair comparison.

To measure performance, standard evaluation metrics including accuracy, precision, recall, and F1-score were used. Accuracy represents the overall correctness of the model, while precision indicates the proportion of correctly identified bullying instances among all predicted positives. Recall measures the model's ability to detect all actual bullying instances, which is particularly important in cyberbullying detection tasks. F1-score provides a balance between precision and recall and is considered the most reliable metric for imbalanced datasets.

The classical machine learning models, including Support Vector Machine (SVM) and Logistic Regression, were implemented using TF-IDF features. These models provided a strong baseline, achieving moderate accuracy but showing limitations in handling contextual and code-mixed language patterns. The SVM model achieved an accuracy of 79.2%, precision of 0.81, recall of 0.72, and F1-score of 0.76, while Logistic Regression showed slightly lower performance with an accuracy of 78.5% and F1-score of 0.74.

The deep learning model, Bidirectional Long Short-Term Memory (Bi-LSTM), demonstrated improved performance by capturing sequential dependencies and contextual relationships in text. The Bi-LSTM model achieved an accuracy of 84.1%, precision of 0.83, recall of 0.81, and F1-score of 0.82, indicating better handling of linguistic variations compared to classical models.

The transformer-based approach using MuRIL significantly outperformed both classical and deep learning models. Due to its pre-training on Indian languages and transliterated text, MuRIL effectively handled Hinglish and code-mixed inputs. The fine-tuned MuRIL model achieved an accuracy of 92.4%, precision of 0.91, recall of 0.93, and F1-score of 0.92, making it the best-performing model in this study.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Model	Accuracy	Precision	Recall	F1-Score
SVM (TF-IDF)	79.2%	0.81	0.72	0.76
Logistic Regression	78.5%	0.80	0.70	0.74
Bi-LSTM	84.1%	0.83	0.81	0.82
MuRIL (Proposed)	92.4%	0.91	0.93	0.92

VIII. APPLICATIONS

The proposed Guardify system has a wide range of practical applications in real-world scenarios where monitoring and moderating online content is essential. With the increasing use of social media and digital communication platforms, the ability to automatically detect cyberbullying in bilingual and code-mixed text provides significant value across multiple domains.

One of the primary applications is in social media platforms such as Twitter, Facebook, Instagram, and YouTube, where large volumes of user-generated content are posted regularly. Guardify can be integrated into these platforms to automatically detect and flag abusive or harmful content in real time, enabling faster moderation and reducing the spread of cyberbullying.

Another important application is in educational institutions, where online learning platforms, student forums, and messaging systems are widely used. The system can help identify instances of cyberbullying among students, allowing administrators and educators to take timely action and create a safer digital environment.

Guardify can also be applied in online gaming and virtual communities, where toxic behavior and abusive communication are common issues. By monitoring in-game chats and community discussions, the system can help maintain a healthy and respectful user environment.

In the field of cybersecurity and digital governance, the system can support law enforcement agencies and regulatory bodies in monitoring harmful online activities. It can assist in identifying patterns of abusive behavior, hate speech, and harassment, contributing to better policy enforcement and public safety.

Additionally, the system can be integrated into customer support systems and online review platforms to filter abusive language in user feedback, improving the quality of interactions between businesses and customers. It can also be used in content moderation tools for forums and websites to ensure compliance with community guidelines.

Finally, Guardify can be extended to research and analytics applications, where it can be used to study trends in online behavior, sentiment analysis, and the spread of harmful content across different regions and languages. This can provide valuable insights for policymakers, researchers, and organizations working toward safer digital ecosystems.

IX. CONCLUSION AND FUTURE WORK

This research presented Guardify, an intelligent cyberbullying detection system designed to address the challenges of bilingual English–Hinglish social media text. The study highlighted the limitations of traditional monolingual approaches in handling code-mixed language and demonstrated the effectiveness of combining robust preprocessing techniques with advanced machine learning and transformer-based models. Through comparative analysis, it was observed that classical models provide a reliable baseline, while deep learning models improve contextual understanding. However, the transformer-based MuRIL model achieved the best performance due to its ability to capture semantic nuances and handle transliteration in Indian languages effectively. The system demonstrated strong performance across evaluation metrics such as accuracy, precision, recall, and F1-score, making it suitable for real-world content moderation applications.

The proposed methodology successfully addressed key challenges such as noise in social media text, inconsistent spelling, emoji usage, and code-mixing. By incorporating transliteration normalization and contextual embeddings, the system improved detection accuracy in complex linguistic environments. The modular architecture of Guardify also ensures scalability and adaptability for deployment across different platforms.

Despite promising results, there are several directions for future improvement. One major limitation is the difficulty in detecting sarcasm, irony, and context-dependent abuse, which often require deeper semantic and contextual understanding beyond text alone.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Future work can explore the integration of multimodal data such as images, videos, and user behavior patterns to enhance detection capabilities. Additionally, incorporating explainable AI techniques can improve transparency and trust in automated moderation systems by providing interpretable predictions.

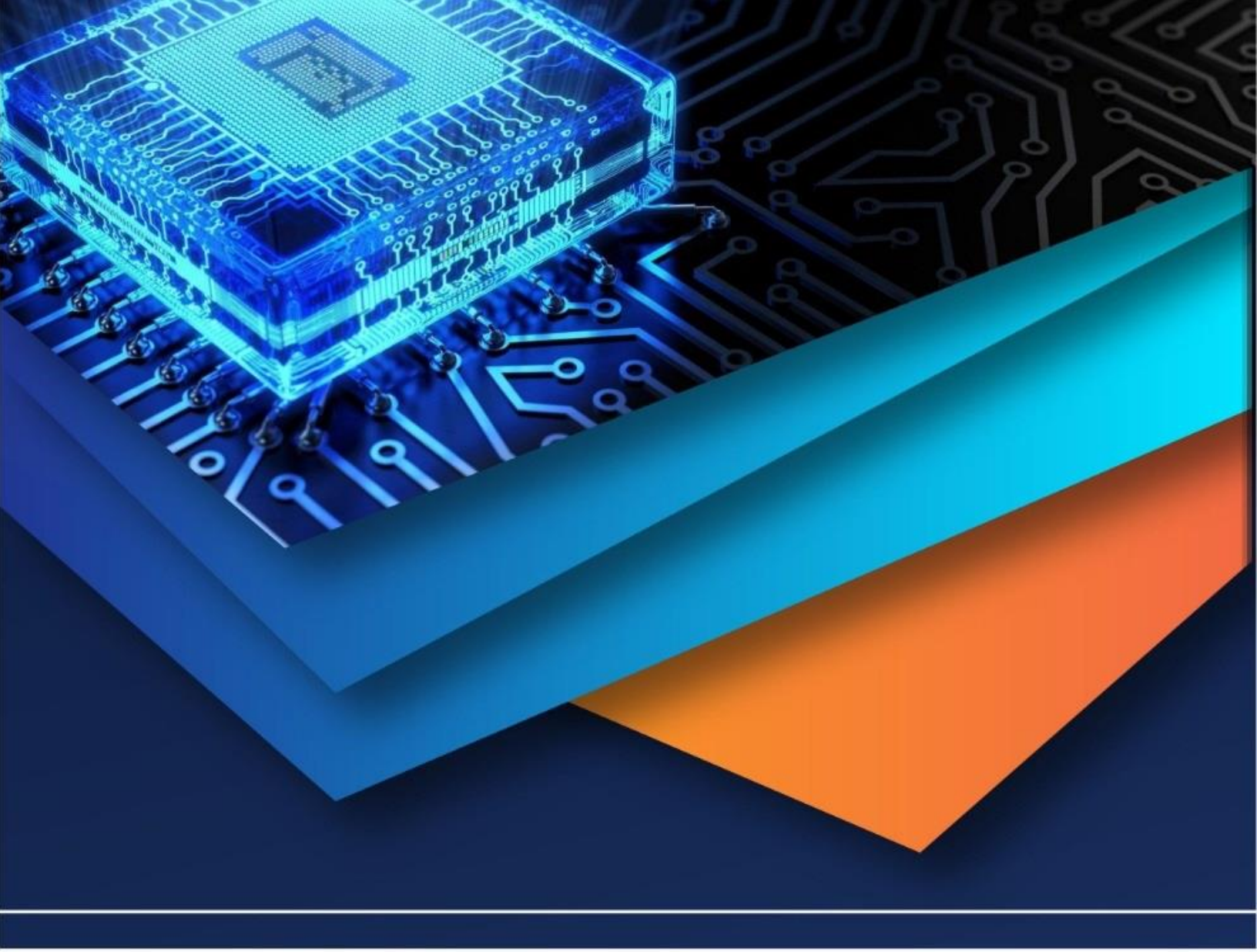
Another area of future research involves continuous model updating to handle evolving slang, emerging abusive patterns, and new linguistic trends in social media. Real-time deployment of the system using APIs and cloud-based infrastructure can further improve scalability and usability. Furthermore, expanding the model to support additional Indian languages and dialects can make the system more inclusive and effective across diverse user communities.

In conclusion, Guardify provides a robust and scalable solution for cyberbullying detection in code-mixed bilingual text, contributing to safer and more responsible digital communication environments. Future enhancements will focus on improving contextual understanding, expanding language coverage, and enabling real-time intelligent moderation systems.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” Proc. ICWSM, 2017.
[2] P. Kumar, “Multilingual Representations for Indian Languages (MuRIL) for Code-Mixed Text Classification,” arXiv preprint arXiv:2103.10730, 2021.
[3] S. Vaswani et al., “Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019.
[5] B. Bohra, D. Vijay, V. Singh, S. Akhtar, and M. Shrivastava, “A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection,” in Proc. Workshop on Abusive Language Online, 2018.
[6] T. Mandl et al., “Overview of the HASOC Track at FIRE 2021: Hate Speech and Offensive Content Identification,” in Proc. FIRE, 2021.
[7] T. Mikolov et al., “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.
[8] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” in Proc. EMNLP, 2014.
[9] K. Cho et al., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” arXiv preprint arXiv:1406.1078, 2014.
[10] Y. Liu, “Fine-Tune BERT for Extractive Summarization,” arXiv preprint arXiv:1903.10318, 2019.
[11] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in Proc. EMNLP, 2004.
[12] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI Whisper, 2022.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)