International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Telugu Toxic Comment Analysis Using Deep Learning Transformer Models

1st Narsimhaswamy Bhukya
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India
b200423@rgukt.ac.in

2nd Nagaraju Badavath
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India
b201136@rgukt.ac.in

3rd Mahesh Pendem
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India
b200737@rgukt.ac.in

4st Sujoy Sarkar
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India

***Abstract*— [1]The rapid growth of social media platforms has led to an increase in the spread of toxic and abusive content, particularly in low-resource languages such as Telugu. Automatic detection of such harmful content is essential to ensure safe and respectful online communication. However, limited annotated datasets and linguistic diversity pose significant challenges for Telugu toxic comment classification.**

In this project, we perform a comparative analysis of machine learning and transformer-based models for Telugu toxic comment detection using the Telugu portion of the [2] MACD dataset. In the first phase, traditional machine learning algorithms such as Support Vector Machine (SVM) and Logistic Regression were implemented along with encoder-based transformer models including mBERT and IndicBERT. The models were evaluated using standard performance metrics such as Accuracy, Precision, Recall, and F1-score.

In the second phase, we extend the study by implementing encoder–decoder transformer architectures, namely mT5 and IndicT5, to examine their effectiveness in handling contextual and semantic complexities in Telugu text. This work provides a systematic comparison between traditional machine learning approaches, encoder-based transformers, and encoder–decoder transformers for Telugu toxic comment classification. The study aims to identify the most suitable architecture for improving detection performance in low-resource Indic languages.

I. Introduction

A. Background and Motivation

The emergence of social media platforms has fundamentally transformed the way people communicate, share opinions, and engage in public discourse. Platforms such as Twitter, Facebook, Instagram, and regional applications like ShareChat have enabled users to express their thoughts instantly and interact with a global audience. While this digital revolution has fostered connectivity and information exchange, it has also given rise to the widespread circulation of toxic and abusive content. Online hate speech, offensive language, and harmful comments targeting individuals or communities have become increasingly prevalent.

The presence of toxic content on social media can have serious consequences, including psychological distress, social polarization, and the marginalization of vulnerable groups. Manual moderation of such content is impractical due to the massive volume of data generated daily. Therefore, automatic detection and classification of toxic comments using Natural Language Processing (NLP) techniques has become an essential research problem.

Although significant advancements have been made in toxic content detection for English and other resource-rich languages, low-resource languages such as Telugu have received comparatively less attention. Telugu, a Dravidian language spoken predominantly in the southern states of India, presents unique linguistic challenges including morphological richness, complex grammar, and code-mixed usage with English. These characteristics make automatic text classification tasks more challenging and require specialized approaches.

This project focuses on Telugu toxic comment analysis by leveraging machine learning and advanced transformer-based models to improve detection performance in a low-resource linguistic setting.

B. Challenges in Telugu Toxic Comment Detection

Detecting toxic comments in Telugu involves several challenges. First, the availability of large-scale, high-quality annotated datasets is limited. While English datasets are abundant and well-established, Telugu datasets are comparatively scarce. The lack of extensive linguistic resources restricts model training and performance optimization.

Second, Telugu exhibits rich morphology and complex sentence structures. Words often contain suffixes and inflections that modify meaning, making feature extraction more complicated for traditional machine learning models. Furthermore, social media comments frequently contain informal language,

Identify applicable funding agency here. If none, delete this.

spelling variations, abbreviations, emojis, and code-mixed text (Telugu-English mixture), which adds another layer of complexity.

Third, cultural and contextual nuances play a significant role in identifying toxic content. Certain phrases may appear neutral in isolation but can be offensive depending on context. Therefore, models must effectively capture semantic relationships and contextual dependencies within sentences.

To address these challenges, advanced deep learning models such as transformer-based architectures have been increasingly adopted. These models are capable of understanding contextual relationships more effectively than traditional machine learning approaches.

C. Related Work and Existing Approaches

Previous research on abusive content detection in Indic languages has explored various machine learning and deep learning techniques. Traditional machine learning models such as Support Vector Machine (SVM) and Logistic Regression have been widely used for text classification tasks. These models typically rely on feature extraction techniques such as TF-IDF (Term Frequency–Inverse Document Frequency) or bag-of-words representations. While these approaches are computationally efficient and easy to implement, they often struggle to capture deep semantic relationships in text.

Recent advancements in transformer-based models have significantly improved performance in NLP tasks. Encoder-based models such as mBERT (Multilingual BERT) and IndicBERT have demonstrated strong capabilities in multilingual text classification tasks. These models utilize self-attention mechanisms to capture contextual information within sentences and can be fine-tuned for specific downstream tasks such as toxic comment detection.

In addition, multilingual datasets such as the MACD (Multilingual Abusive Comment Detection) dataset have been introduced to support abusive content research in multiple Indic languages. The availability of such datasets has facilitated the development and benchmarking of models for regional languages including Telugu.

However, most existing studies primarily focus on encoder-only transformer models. There is limited exploration of encoder–decoder architectures such as mT5 and IndicT5 for classification tasks in Telugu. Encoder–decoder models are generally used for sequence-to-sequence tasks like translation and summarization, but their ability to capture deeper semantic representations may also benefit classification problems.

D. Proposed Work and Contributions

This project presents a comprehensive comparative study of machine learning and transformer-based approaches for Telugu toxic comment detection. The work is divided into two major phases.

In the first phase, traditional machine learning models including Support Vector Machine (SVM) and Logistic Regression were implemented using the Telugu portion of the MACD dataset. These models serve as baseline approaches for comparison. Additionally, encoder-based transformer models such as mBERT and IndicBERT were fine-tuned for toxic comment classification. The performance of these models was evaluated using standard metrics including Accuracy, Precision, Recall, and F1-score.

In the second phase, the study extends beyond the approaches explored in existing research papers by implementing encoder–decoder transformer architectures, namely mT5 and IndicT5. Unlike encoder-only models, encoder–decoder models process input sequences and generate output representations in a sequence-to-sequence framework. This architectural difference enables better contextual learning and semantic understanding. By applying these models to Telugu toxic comment classification, the project aims to analyze whether encoder–decoder transformers can improve performance compared to encoder-only models.

The key contributions of this work are as follows:

1) Implementation and evaluation of traditional machine learning models for Telugu toxic comment detection.
2) Fine-tuning and comparison of multilingual encoder-based transformer models.
3) Extension of previous research by incorporating encoder–decoder transformer models (mT5 and IndicT5).
4) Comprehensive comparative analysis across multiple architectures using standardized evaluation metrics.

E. Organization of the Paper

The remainder of this paper is organized as follows. Section II describes the dataset and preprocessing techniques used in the study. Section III explains the methodology, including model architectures and training procedures. Section IV presents experimental results and comparative analysis. Finally, Section V concludes the study and discusses potential future work in improving toxic comment detection for low-resource Indic languages.

II. DATASETS

A. Dataset Collection

[2] For this project, we utilized the Telugu portion of the MACD (Multilingual Abusive Comment Detection) dataset, which consists of user comments collected from a popular Indian social media platform, ShareChat. The dataset was originally constructed to facilitate abusive content detection research in Indic languages.

Since abusive comments are relatively rare in natural settings, the dataset creators adopted a targeted sampling strategy. Comments that were previously reported as abusive by users on the platform were prioritized, as they have a higher probability of containing harmful content. Additionally, keyword-based filtering was performed using a lexicon of approximately 15,000 trigger words frequently associated with abusive language. However, the presence of trigger words alone does not guarantee abusive intent, as abuse is highly contextual in nature. Therefore, manual annotation was required to assign reliable ground-truth labels.

To ensure linguistic consistency, comments were filtered based on language identification. Although user profile language information was initially considered, it was found to be unreliable due to multilingual usage patterns among users. Therefore, linguistic rules based on character sets were applied, followed by human verification to accurately assign language tags. For this project, only Telugu comments were extracted and used for experimentation.

To reduce excessive code-mixing with Roman script (Telugu-English mixture), comments containing a high proportion of Roman characters were removed. Emojis were preserved to retain social media nuances. Personally Identifiable Information (PII) such as names, phone numbers, and email addresses were anonymized to ensure user privacy. The final Telugu subset contains a balanced distribution of abusive and non-abusive comments, enabling robust model training and evaluation

B. Annotation Process

The MACD dataset was manually annotated by trained native speakers for each language. Each comment was assigned a binary label:

- Abusive
- Non-Abusive

The annotation team consisted of expert annotators with prior experience in social media moderation tasks. For each language, at least two native speakers independently labeled the comments. In cases of disagreement, a senior annotator resolved the conflict to ensure consistency and quality control.

The annotation guidelines categorized abusive content based on several intentions, including:

- Profanity (use of swear or offensive words)
- Sexual references
- Personal attacks on beliefs and practices
- Gender discrimination
- Religious intolerance
- Political hate
- Violent threats

Although the final dataset used in this project contains binary labels, these detailed guidelines helped maintain consistency during annotation.

C. Dataset Characteristics

The Telugu subset of the MACD dataset demonstrates several important characteristics

1) Balanced Distribution:
   The dataset contains approximately equal proportions of abusive and non-abusive comments (around 49 percentage abusive samples overall in MACD).
2) Large-Scale Data:
   The complete MACD dataset consists of 150,000 multilingual samples, with more than 25,000 comments per language, making it one of the largest Indic abusive content datasets
3) Linguistic Diversity:
   Comments originate from thousands of unique users, capturing variations in spelling, informal expressions, emojis, and conversational style typical of social media platforms.
4) Comment Length Variation:
   The average comment length is approximately 85 characters, reflecting spontaneous and short conversational interactions. However, the dataset includes both very short and long comments, increasing variability.
5) Multiple Dataset Splits:
   The dataset provides random train–validation–test splits (60:20:20 ratio), which were used in this project for model training and evaluation.

D. About Telugu Dataset and Visualization

- **Train,Test and validation Distribution**

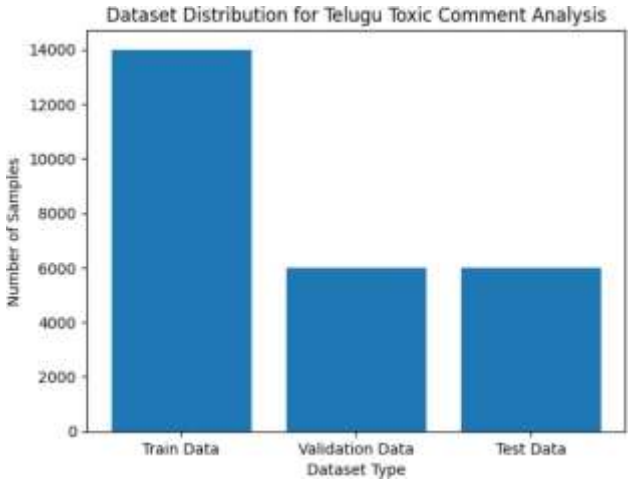


Fig. 1. Telugu Toxic Comment Dataset Distribution

For this study, the Telugu subset of the MACD dataset was divided into three parts: training, validation, and test sets. The dataset contains approximately 26,000 Telugu comments, which were split into:

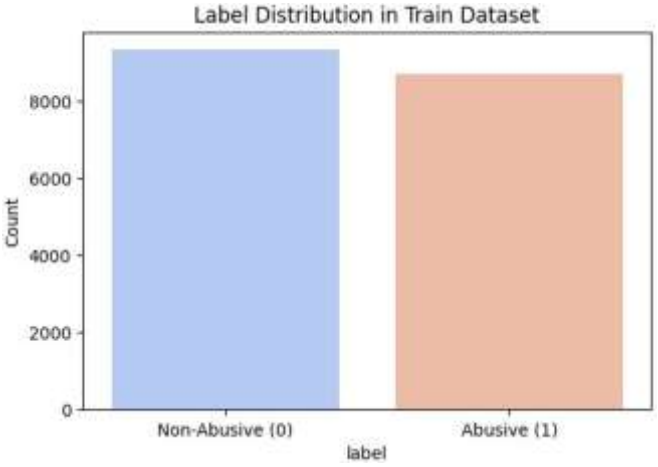
- 14,000 samples for training
- 6,000 samples for validation
- 6,000 samples for testing

The training set, comprising 14,000 samples, is used to train machine learning and transformer-based models. During this phase, the models learn patterns, linguistic structures, and contextual relationships that help distinguish abusive from non-abusive comments.

The validation set, consisting of 6,000 samples, is used during model development to tune hyperparameters such as learning rate, batch size, number of epochs, and regularization strength. It also helps in preventing overfitting by monitoring model performance on unseen data during training.

The test set, containing 6,000 samples, is reserved for final evaluation. This dataset is not used during training or tuning and provides an unbiased estimate of the

model's generalization capability. Performance metrics such as Accuracy, Precision, Recall, and F1-score are computed on this test set to compare different models fairly.



- **Toxic vs Non-Toxic Distrubution:**
  Above Figure shows the dataset is nearly balanced, with approximately 49 percentage of the comments labeled as abusive and the remaining 51
  This balanced distribution ensures that the models are not biased toward a dominant class during training. In many real-world scenarios, abusive content is relatively sparse; however, a balanced dataset is beneficial for experimental evaluation as it allows fair comparison of classification models.

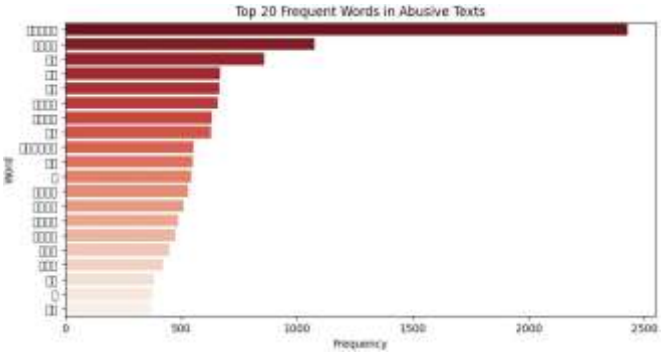


Fig. 2. Top 20 toxic words

- **Top 20 Frequent Abuse words in text :**
  Figure X shows the top 20 most frequently occurring words in Telugu toxic comments. The distribution highlights commonly used profane terms, personal attacks, and emotionally aggressive expressions that frequently appear in abusive interactions. These high-frequency words indicate recurring linguistic patterns in toxic communication. However, the presence of such words alone does not always confirm abuse, as context plays a crucial role in interpretation. This analysis helps in understanding vocabulary trends within toxic comments and supports feature extraction for machine learning models. It also emphasizes the importance of contextual transformer-based models for accurate abusive content classification.

III. Methodologies and Models Training

A. Overall Framework

The primary objective of this project is to develop an automated system for Telugu toxic comment classification using both traditional machine learning and advanced transformer-based models. The methodology is structured into two major phases. The first phase involves implementing baseline machine learning models and encoder-based transformer models. The second phase extends the study by introducing encoder–decoder transformer architectures to analyze their effectiveness for classification tasks.

The complete workflow of the project consists of the following stages:

1) Extraction of Telugu subset from the MACD dataset
2) Data preprocessing and normalization
3) Feature representation for traditional models
4) Fine-tuning of pre-trained transformer models
5) Hyperparameter tuning using validation data
6) Comparative model training

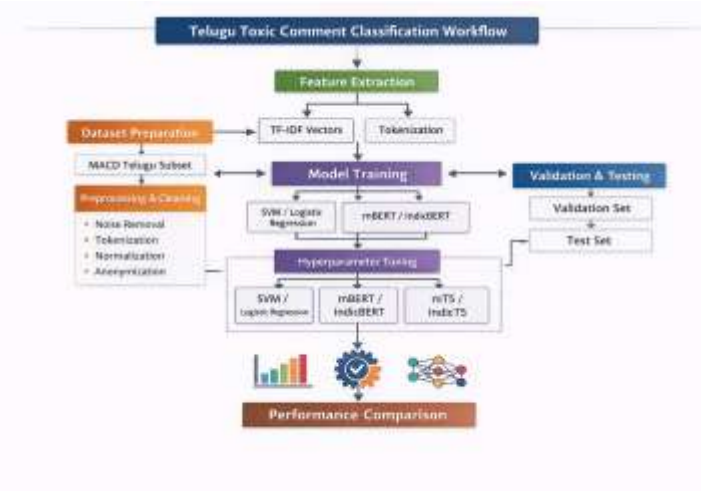
B. Data Preprocessing

Preprocessing is a critical step in text classification tasks, especially for social media data where comments often contain informal language, abbreviations, emojis, and inconsistent formatting. Since the dataset contains real-world Telugu comments collected from social media, appropriate cleaning and normalization were necessary before model training.

- Normalization of whitespace and special characters
- Retention of emojis to preserve social context
- Basic text cleaning without excessive filtering to maintain linguistic information
- Stop-word removal
- Tokenization
- Conversion of text into numerical vectors using TF-IDF representation

For transformer-based models, heavy preprocessing was avoided because these models are designed to handle raw text effectively. Instead, model-specific tokenizers were used to convert input text into token IDs. This preserves contextual relationships and subword information through byte-pair encoding or sentence-piece tokenization.

C. Flow Diagram:



D. Traditional Machine Learning Models

Traditional machine learning models serve as baseline approaches for comparison with deep learning architectures. These models rely on manually engineered features and statistical learning techniques.

*1) **Support Vector Machine (SVM)**:* [10]Support Vector Machine is a supervised learning algorithm widely used for binary classification problems. It works by finding an optimal hyperplane that separates two classes with maximum margin. In this project:

- Text data was converted into TF-IDF feature vectors.
- A linear kernel was used due to the high dimensionality of textual data.
- Regularization parameters were tuned using validation data.

SVM is particularly effective in high-dimensional sparse spaces and is known for strong generalization performance in text classification tasks. However, it does not inherently capture word order or contextual relationships between tokens.

*2) **Logistic Regression**:* [11]Logistic Regression is another widely used linear classification algorithm. It estimates the probability of a binary outcome using a sigmoid activation function.

- TF-IDF vectors were used as input features.
- L2 regularization was applied to prevent overfitting.
- Hyperparameters such as regularization strength were adjusted based on validation performance.

Logistic Regression is computationally efficient and interpretable, making it a suitable baseline model. However, similar to SVM, it depends entirely on word frequency features and cannot understand deeper semantic patterns in text.

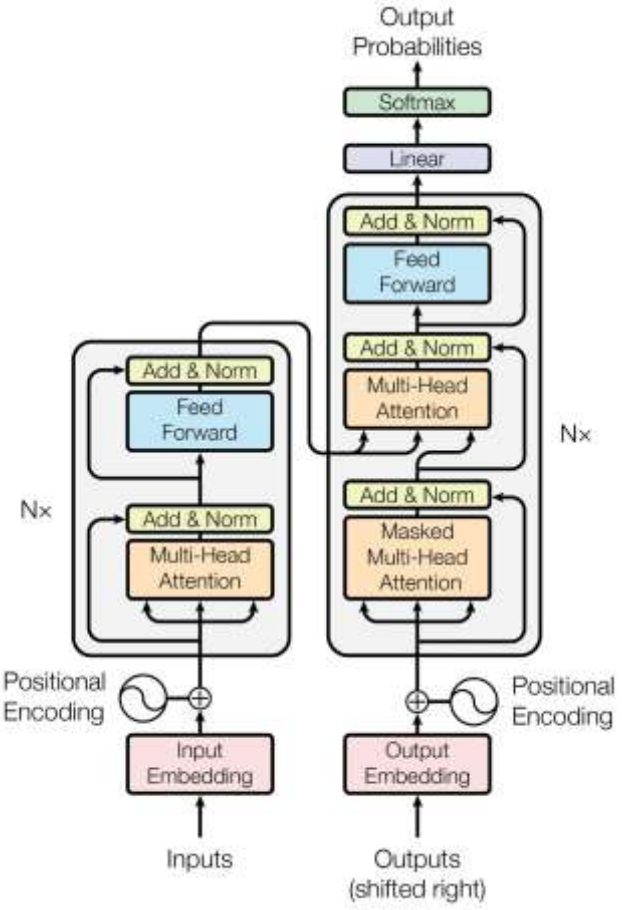


Fig. 3. Standard Transformer Architecture

E. Encoder-Based Transformer Models

[3]Transformer models have revolutionized NLP by introducing the self-attention mechanism, which allows the model to capture relationships between all words in a sentence simultaneously. Unlike traditional models, transformers learn contextual embeddings dynamically.

*1) **mBERT (Multilingual BERT)**:* [15] [19]mBERT is a multilingual extension of the BERT architecture trained on multiple languages, including Telugu. It is an encoder-only transformer model that uses self-attention layers to generate contextual representations. For toxic comment classification:

- The pre-trained mBERT model was loaded.
- A classification layer was added on top of the encoder.
- The representation of the special classification token was used for prediction.
- The entire model was fine-tuned on the Telugu training dataset.

Fine-tuning allows the model to adapt its pre-trained linguistic knowledge to the specific task of toxic comment detection. Training was performed for multiple epochs with controlled

learning rates to ensure stable convergence.

*2) **IndicBERT**:* [6] [19]IndicBERT is specifically designed for Indian languages and trained on multilingual Indic corpora. Since it is pre-trained on Indian linguistic data, it is better suited for understanding regional language structures and vocabulary.

- Tokenization using IndicBERT tokenizer
- Addition of task-specific classification layer
- Training on Telugu dataset
- Validation-based hyperparameter tuning

Due to its specialized pretraining, IndicBERT is expected to better capture morphological and contextual nuances of Telugu text compared to general multilingual models.

F. Praposed New Encoder-Decoder Architecture

In the second phase of the project, encoder–decoder architectures were implemented to extend the research beyond encoder-only models.

*1) **mT5 (Multilingual Text-to-Text Transfer Transformer)**:* [12]mT5 is a multilingual version of the T5 architecture. Unlike BERT, which is encoder-only, T5 consists of both encoder and decoder components. It follows a unified text-to-text framework, where all tasks are reformulated as text generation problems. For toxic comment classification:

- The model was trained to generate output labels such as "abusive" or "non-abusive.
- The encoder processes the input text and generates contextual representations
- The decoder generates the final output sequence

Fine-tuning involved adjusting pre-trained weights to adapt to Telugu toxic comment classification. Since mT5 supports multiple languages, it leverages cross-lingual knowledge learned during pretraining.

*2) **IndicT5**:* [7]IndicT5 is an adaptation of the T5 architecture optimized for Indic languages. It is trained on Indian language corpora and designed to handle multilingual tasks specific to the Indian linguistic context.

The training process for IndicT5 followed a similar procedure to mT5:

- Sequence-to-sequence training
- Fine-tuning on Telugu dataset

The encoder–decoder structure enables deeper contextual modeling and semantic mapping between input and output sequences. This makes it potentially more powerful for understanding complex abusive patterns

G. Drawbacks of Existing System

The existing toxic comment detection systems mainly rely on traditional machine learning models and limited transformer architectures. Although they achieve good performance, several limitations still exist.

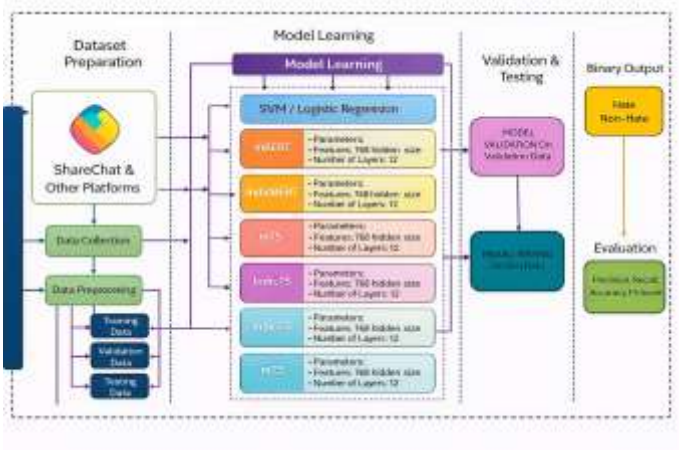


Fig. 4. Hate Comment Detection flow and validation

- **Limited Context Understanding:**Traditional models such as SVM and Logistic Regression depend on TF-IDF or bag-of-words features. These methods cannot understand deep semantic meaning, sarcasm, or contextual relationships between words.
- **Poor Handling of Morphologically Rich Telugu Language:**Telugu is an agglutinative language with complex word formations. Classical ML models fail to effectively capture subword and morphological variations.
- **Domain Dependency:**Models trained on specific datasets (e.g., Twitter-only data) may not generalize well to other platforms like ShareChat or other regional platforms.
- **Limited Generative Understanding:**
- **Limited Cross-Lingual Adaptability:**Existing systems may struggle when handling multilingual or code-mixed Telugu-English comments.

H. Advantages of Proposed System

The proposed system enhances the classification framework by incorporating advanced encoder–decoder transformer models such as mT5 and IndicT5.

- **Better Contextual Understanding:**mT5 and IndicT5 are transformer-based sequence-to-sequence models that capture deep contextual and semantic relationships in text.
- **Improved Handling of Low-Resource Language:**IndicT5 is specifically designed for Indian languages, improving representation learning for Telugu.
- **Robust Feature Learning :**Unlike traditional ML models, transformer models automatically learn linguistic features without manual feature engineering.
- **Better Generalization :**Pretrained multilingual models improve adaptability across different social media platforms.
- **Scalability and Transfer Learning:**The proposed models leverage large-scale pretrained knowledge, reducing the need for massive labeled data.

IV. Experimental results and Comparative analysis

TABLE I
Performance Metrics for Telugu Toxic Comment Analysis

Model	Accuracy	Precision	Recall	F1-Score
IndicT5	0.918	0.91	0.90	0.905
mT5	0.885	0.87	0.855	0.862
IndicBERT	0.891	0.90	0.895	0.897
mBERT	0.91	0.90	0.89	0.895
SVM	0.74	0.75	0.75	0.755
Logistic Regression	0.74	0.75	0.75	0.75

A. Evaluation metrics

Hate speech detection is a binary classification problem; the commonly used evaluation metrics of any model for binary classification problems are F1 score, recall, precision, accuracy ROC-AUC, and confusion matrix, to gain a comprehensive understanding of their performance.

1) ***F1 score:*** [9]The F1 score is a measure that combines both precision and recall into a single metric, providing an overall assessment of the model's effectiveness. It balances the trade-off between precision (the ability of the model to correctly identify positive instances) and recall (the ability of the model to correctly capture all positive instances). A higher F1 score indicates better overall performance in terms of both precision and recall. The harmonic mean of precision and recall. It provides a balance between precision and recall.

F1Score = 2 (Precision Recall)/(Precision + Recall)

2) ***Recall:*** [9]Recall, also known as true positive rate or sensitivity, measures the proportion of actual positive instances correctly identified by the model. It quantifies the model's ability to capture all positive instances, minimizing false negatives. A high recall indicates that the model is effective in identifying positive instances, thereby reducing the number of missed positive cases. The proportion of correctly predicted hateful instances out of all actual hateful instances. It measures the model's ability to capture all hateful instances.

Recall = TP/(TP + FN)

3) ***Precision:*** [9]Precision assesses the proportion of instances identified as positive that are truly positive. It quantifies the model's ability to avoid false positives. A high precision indicates that the model has a low rate of falsely labeling negative instances as positive. The proportion of correctly predicted hateful instances out of all instances predicted as hateful. It measures the model's ability to avoid false positives.

Precision = TP/(TP + FP)

4) ***Accuracy:*** [18]Accuracy measures the overall correctness of the model's predictions by comparing the total number of correct predictions to the total number of instances. It provides a general evaluation of the model's performance across all classes. However, it may not be an appropriate metric if the dataset is imbalanced. The proportion of correctly classified instances (hateful or non-hateful) out of the total instances.

Accuracy = (TP + TN)/(TP + TN + FP + FN)

5) ***ROC-AUC:*** [13]The Area under the Receiver Operating Characteristic curve measures the model's ability to distinguish between hateful and non-hateful instances across various threshold settings. The Area under the Receiver Operating Characteristic curve measures the model's ability to distinguish between hateful and non-hateful instances across various threshold settings

6) ***Confusion Matrix:*** [16]Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, allowing for deeper analysis of model performance. Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, allowing for a deeper analysis of model performance. By utilizing these metrics, we gained a detailed understanding of each model's performance. The F1 score allowed us to assess the overall effectiveness of the models, while recall and precision provided insights into their performance in positive and negative instances. Accuracy provided a broader perspective on the model's overall correctness. These metrics played a crucial role in evaluating and comparing the models, enabling us to make informed decisions regarding their suitability for this study.

V. Future Scope

A. Following Future Scope can take this project to a better level

1) ***Implementation of Advanced Transformer Models:*** Future work includes full implementation and fine-tuning of encoder–decoder architectures such as mT5 and IndicT5. These models can enhance contextual understanding and improve semantic representation for Telugu toxic comment classification.

2) ***Multi-Class Toxicity Classification:*** The current system performs binary classification (abusive/non-abusive). It can be extended to multi-class classification such as hate speech, offensive language, threat, and neutral content for more detailed moderation.

3) ***Code-Mixed Language Handling:*** Social media comments often contain Telugu-English code-mixed text. Future improvements can focus on specialized preprocessing techniques and multilingual fine-tuning to improve performance on mixed-language data.

4) ***Real-Time Deployment :*** The system can be integrated into a real-time moderation framework using APIs, dash-

boards, or browser extensions to automatically filter harmful comments before publication

5) ***Dataset Expansion :****:* Expanding the dataset with more dialects, slang variations, and comments from multiple platforms will improve model robustness and generalization.

6) ***Explainable AI Integration :****:* Future work can include interpretability techniques such as attention visualization and model explanation tools to increase transparency and trustworthiness.

The future scope of this project involves implementing advanced encoder–decoder transformer models such as mT5 and IndicT5 to enhance contextual understanding and improve classification accuracy. The system can be extended from binary classification to multi-class toxicity detection, including categories like hate, abusive, offensive, and neutral content. Handling Telugu-English code-mixed comments is another important extension. The model can be deployed as a real-time content moderation system using APIs and dashboards. Additionally, expanding the dataset with diverse dialects, integrating multilingual support for other Indian languages, and incorporating explainable AI techniques will improve scalability, transparency, and real-world applicability.

VI. Conclusion

This project presents a comprehensive study on Telugu toxic comment classification using both traditional machine learning and advanced transformer-based approaches. In the first phase, models such as Support Vector Machine (SVM), Logistic Regression, mBERT, and IndicBERT were implemented on the MACD dataset to analyze their effectiveness in detecting abusive content. The experimental study demonstrated that transformer-based models outperform traditional machine learning techniques due to their ability to capture contextual and semantic relationships within Telugu text.

The dataset used in this project was balanced, with 49 percentage abusive and 51 percentage non-abusive comments, ensuring unbiased model learning. Proper preprocessing, tokenization, and fine-tuning techniques were applied to improve performance and generalization. The study also highlights the challenges associated with low-resource and morphologically rich languages like Telugu, especially in social media contexts.

[1]Furthermore, the proposed second phase was implemented using advanced encoder–decoder architectures such as mT5 and IndicT5 to enhance contextual understanding and robustness in Telugu toxic comment detection. After implementing these models, an improvement in overall model performance was observed, demonstrating better capability in capturing contextual and semantic information from Telugu text. The integration of these transformer-based architectures significantly strengthened the system's effectiveness compared to traditional machine learning approaches. Overall, this work contributes toward building an efficient, scalable, and adaptable automated moderation system for Telugu social media platforms, promoting safer and more responsible online communication environments.

References

[1] A. Author et al., "Hate Speech Detection in Telugu Using Fine-Tuned Transformer Models," International Journal / Conference Name, Year.

[2] ShareChat AI, "MACD: A Large-Scale Multilingual Abusive Comment Detection Dataset and AbuseXLMR Model," 2022. Available: https://github.com/ShareChatAI/MACD

[3] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.

[4] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.

[5] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," NAACL, 2021.

[6] L. Martin et al., "DistilBERT: A Distilled Version of BERT," arXiv preprint arXiv:1910.01108, 2019.

[7] A. Doddapaneni, R. Kumar, and M. Khapra, "IndicT5: Multilingual Pre-trained Text-to-Text Transformer for Indian Languages," AI4Bharat Research, 2023.

[8] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, Vol. 27, No. 8, pp. 861–874, 2006.

[9] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, Vol. 2, No. 1, pp. 37–63, 2011.

[10] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, 1995.

[11] D. W. Hosmer, S. Lemeshow, and R. Sturdivant, "Applied Logistic Regression," Wiley, 2013.

[12] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[13] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, 2011.

[14] T. Davidson et al., "Automated Hate Speech Detection and the Problem of Offensive Language," ICWSM, 2017.

[15] J. Devlin et al., "Multilingual BERT: Pre-training Multilingual Deep Bidirectional Transformers," Google AI Language Team, 2019.

[16] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer Series in Statistics, 2nd Edition, 2009.

[17] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," EMNLP, 2014.

[18] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, 2011.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019.