UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

CYBERBULLYING DETECTION ON SOCIAL MEDIA TEXTUAL CONTENTS USING DEEP LEARNING

Dr. S. Komal Kour[1]
Dept. Computer Science and Engineering
Vasavi College of Engineering (Autonomous) Ibrahimbagh,Hyderabad
Email: komalkour@staff.vce.ac.in

Muthyala Rakesh[2]
Dept. Computer science and engineering
Vasavi College of Engineering (Autonomous) Ibrahimbagh,Hyderabad
Email: muthyalarakesh999@gmail.com

Abstract- Cyberbullying has emerged as a significant concern across social media platforms, given its profound psychological and societal impact. Traditional moderation strategies—such as reporting or blocking users—are largely manual and often fall short in effectively mitigating harmful content. To address this gap, we present an automated cyberbullying detection system powered by a hybrid deep learning model that integrates XLNet and RoBERTa architectures. This combination harnesses the contextual understanding capabilities of Transformer-based language models, enabling more accurate and reliable identification of abusive language. The model is trained on a curated, balanced dataset that includes various forms of cyberbullying, such as those targeting gender or sexual orientation, ethnicity or race, and religion, along with neutral (non-cyberbullying) content.

Our hybrid approach significantly outperforms traditional machine learning techniques and standalone deep learning models, reaching a validation accuracy of approximately **96.7%**. Unlike earlier methods that require extensive manual feature extraction and data preprocessing, our solution benefits from transfer learning to simplify the pipeline while boosting performance. For ease of use, the model is deployed through a Gradio-based interface, enabling real-time user interaction and content evaluation. The results underscore the potential of hybrid Transformer architectures in building scalable, efficient, and accessible tools for tackling online abuse.

Keywords

Cyberbullying Detection, XLNet, RoBERTa, Hybrid Model, Transformer Architecture, Deep Learning, Transfer Learning, social media, Natural Language Processing, Realtime Prediction.

I.INTRODUCTION

Cyber bullying has become a pressing issue on social media due to its deep psychological and social repercussions. Existing moderation methods, such as user reports or blocking mechanisms, are mostly manual and often ineffective in addressing the scale and subtlety of online abuse. To overcome these limitations, we propose an automated cyberbullying detection system built on a hybrid deep learning framework that combines the strengths of XLNet and RoBERTa. By leveraging the contextual capabilities of advanced Transformer-based

models, our system delivers improved accuracy in detecting harmful and abusive content.

The model is trained on a well-balanced and preprocessed dataset that includes diverse categories of cyberbullying, such as those based on gender identity, sexual orientation, ethnicity, race, and religion, as well as non-offensive content. Experimental results demonstrate that the hybrid model outperforms both traditional machine learning algorithms and standalone deep learning architectures, achieving a validation accuracy of approximately 96.7%. Unlike older techniques that depend heavily on manual feature engineering and complex preprocessing, our approach benefits from transfer learning, resulting in faster development and higher efficiency.For practical deployment, the model is integrated into a user-friendly Gradio interface, allowing real-time interaction and predictions. Overall, the results highlight the effectiveness of hybrid Transformer models in building scalable, robust, and accessible tools for mitigating cyberbullying on digital platforms.

II.LITERATURE SURVEY

The landscape of cyberbullying detection has evolved dramatically with advances in Natural Language Processing (NLP) and deep learning. Initial efforts relied on Conventional Machine Learning (CML) techniques such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression. These models typically used handcrafted features like bag-of-words, TF-IDF scores, and sentiment metrics [1]. While they provided a useful starting point for text classification, they often struggled to capture the subtle context and semantics present in online conversations. Attempts to improve these models by incorporating user behavior data, psycholinguistic markers, and emotion-based lexicons helped to an extent, but generalization remained a significant limitation [2].The introduction of deep learning marked a turning point. Neural network architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks began to demonstrate superior performance by learning complex syntactic patterns and sequential dependencies in text [3]. In particular, LSTM models showed strong capabilities in recognizing temporal dynamics in cyberbullying language, which is often informal or coded [4].

However, these models came with their own challenges—they required large datasets, significant computational resources, and long training times.This led to the rise of Transfer Learning and the use of pre-trained language models. Transformer-based architectures like BERT, RoBERTa, and XLNet revolutionized the field by providing contextual word embeddings and employing attention mechanisms to understand bidirectional relationships within text [5]. Studies have consistently shown that fine-tuned versions of BERT outperform both traditional ML models and earlier deep learning methods across various cyberbullying datasets [6]. Lightweight alternatives such as DistilBERT and Electra-small emerged to reduce inference time and computational overhead while still maintaining strong performance [7]. Notably, DistilBERT has delivered high F1 scores in cyberbullying classification, especially when enriched with additional features like toxicity levels, sentiment polarity, and psychological language patterns [8].More recent research has shifted toward hybrid and ensemble models that combine the strengths of multiple transformer architectures.

For instance, blending RoBERTa's deep contextual insight with XLNet's permutation-based learning has shown promising results, particularly in improving classification accuracy across categories such as race, gender, religion, and neutral content [9]. These hybrid solutions not only boost accuracy but also enhance robustness in handling diverse online conversations. On the deployment front, interactive tools like Gradio have made it easier to integrate these models into real-world applications, offering

user-friendly interfaces for live prediction and analysis [10].Overall, the field has progressed from simple manual feature engineering to sophisticated, context-aware, and scalable detection systems driven by the power of transformers and hybrid deep learning models.

III. PROPOSED METHODOLOGY

The proposed system tackles the rising issue of cyberbullying on social media by utilizing a hybrid deep learning approach that combines the strengths of XLNet and RoBERTa. XLNet's unique permutation-based language modeling helps the system understand a wide range of contextual relationships in text, while RoBERTa contributes deep semantic insight through its powerful bidirectional encoding.

This synergy allows the model to accurately detect different forms of cyberbullying, including abuse based on gender, race, religion, and other sensitive areas.The model is trained on a thoroughly cleaned and balanced dataset, ensuring fair representation across all categories and reducing potential bias.

By fine-tuning pre-trained transformer models, the system avoids the need for complex manual feature extraction while also shortening training times. To make the solution more accessible and practical, the trained model is deployed using a user-friendly Gradio interface. This enables real-time predictions and offers a valuable tool for users and organizations aiming to monitor online behavior and foster safer digital environments.

Dataset

To support the development of an effective cyberbullying detection system, a custom dataset was curated by the authors. This dataset comprises a diverse and balanced collection of social media posts, annotated across multiple categories of cyberbullying. These categories include content targeting gender or sexual orientation, ethnicity or race, religion, and instances of non-cyberbullying. Careful preprocessing steps were undertaken to remove duplicates, noise, and irrelevant text, resulting in a clean, high-quality dataset suitable for training deep learning models.

The balanced distribution of classes ensures that the model learns to detect various forms of cyberbullying without bias toward any particular category. The dataset's scale and diversity make it especially suitable for fine-tuning large transformer-based models like XLNet and RoBERTa, enabling the

development of robust and generalized detection capabilities.

Annotations were conducted in a structured manner to maintain consistency and accuracy across labels. This thorough coverage of real-world cyberbullying contexts enhances the dataset's relevance and applicability to automated content moderation systems.

Figure 1: Sample entry from the dataset used for training. Each row includes a raw social media post (Text) and a binary label (CB_Label), where 1 indicates a cyberbullying instance and 0 denotes non-cyberbullying. The full dataset contains **100,000 records**.

	Text	CB_Label
0	@ChrisWarcraft bro	0
1	RT @Cyn_Santana: Yoooooo RISE N GRIND. THERES ...	0
2	4 Reasons Why Men Joke About Gay Rape http://e...	1
3	.@patrickjbutler Yes, but @EricPickles shd be...	0
4	Posted: Don't Just Stand There: Stop Bullying ...	0
...	...	...
99995	pleas i highly doubt anyone will see this but ...	1
99996	So what was the opposite of a reindeer sweater...	0
99997	Statistically most of the [rattlesnake] bites ...	0
99998	this is why asaf is letting ME in the fake sce...	1
99999	Catching up with #MKR. If society judges these...	1

100000 rows × 2 columns

Figure 1: rows of cyberbullying dataset used for classification

Figure 2 illustrates the architecture of the proposed hybrid deep learning model that integrates XLNet and RoBERTa to enhance cyberbullying detection through combined contextual and semantic understanding.

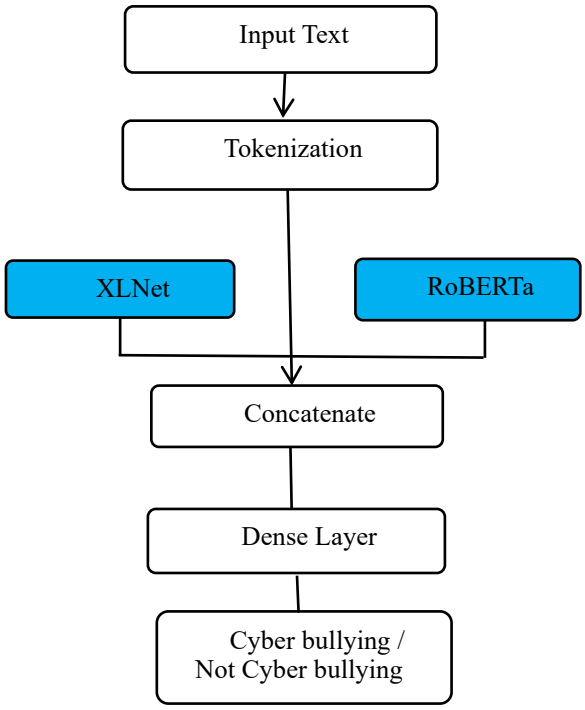


Figure2: Architecture of XLNET+ROBERTA (hybrid model)

XLNet-RoBERTa Based Cyberbullying Detection

The architecture of the proposed hybrid model, illustrated in **Figure 2**, represents a powerful integration of two state-of-the-art transformer-based language models: **XLNet** and **RoBERTa**. This dual-stream configuration combines the unique strengths of both models—XLNet's permutation-based sequence modeling captures diverse contextual dependencies, while RoBERTa provides deep semantic understanding through its robust bidirectional encoding.This combination is particularly effective for analyzing social media content, where language can be informal, nuanced, or intentionally obfuscated. The hybrid model is designed to detect both explicit and subtle forms of cyberbullying across different categories such as gender, race, and religion.The input to the system consists of raw social media text, which is simultaneously processed by the XLNet and RoBERTa pipelines. Each model independently generates rich contextual embeddings, capturing various linguistic features and semantic cues. These embeddings are then **concatenated or merged**, and passed through a series of **fully connected dense layers**, which learn to map the combined features to binary classification labels—indicating whether a post is cyberbullying or not.This hybrid architecture improves **generalization** across diverse datasets, enhances **robustness** to informal and adversarial language, and reduces **false positive rates**, making it suitable for real-time cyberbullying detection in practical applications.

Prediction Interface for Cyberbullying Detection

To enable practical and real-time application, the system features an interactive prediction interface designed to evaluate text inputs on the fly using the trained XLNet–RoBERTa hybrid model. When a user submits a social media post through the interface, the text is first tokenized independently for both XLNet and RoBERTa to maintain compatibility with each model's input requirements.The tokenized representations are then passed through their respective model streams to generate contextual embeddings. These are combined and processed by the model's classification layers, producing raw prediction logits. A softmax activation function is applied to transform these logits into probability scores for each class.The class with the highest confidence score is selected as the final prediction. The system then presents an easy-to-understand output—either **"Cyberbullying"** or **"Not Cyberbullying"**—allowing users or content moderation systems to quickly interpret and act on the result. This interface bridges the gap between advanced NLP research and practical content moderation, making cyberbullying detection both accessible and efficient.

IV.IMPLEMENTATIONAND RESULTS

The visual workflow diagram above illustrates the complete system architecture, showing how raw social media text flows through parallel transformer models to produce accurate cyberbullying detection with automated response capabilities.

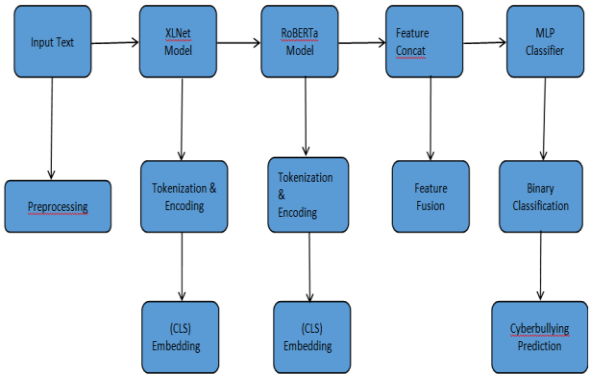


Figure3: Working of the developed system

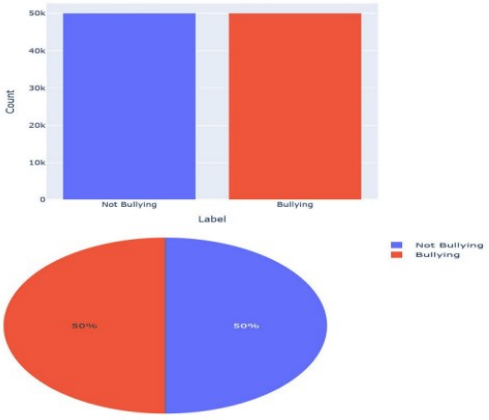
A. Data Preprocessing and Model Training

Data Preprocessing

To ensure the consistency and interpretability of social media text, a comprehensive text normalization process was applied during preprocessing. This step is crucial for minimizing variability in informal language and enhancing the hybrid XLNet–RoBERTa model's ability to learn meaningful patterns. The normalization workflow included converting all text to lowercase, removing punctuation and special characters, standardizing whitespace, eliminating stop words, and applying lemmatization to reduce words to their base forms. These operations help retain semantic meaning while reducing noise and feature sparsity within the dataset. After normalization, the dataset was split into training and testing sets using a stratified 80:20 ratio, ensuring balanced representation of all cyberbullying categories across both subsets. The cleaned text was then tokenized using model-specific tokenizers tailored to XLNet and RoBERTa, allowing the data to be effectively processed by the transformer architectures. The key components and tools used in this preprocessing pipeline, as well as their respective functions, are summarized in data.

Balancing the Data

To address the issue of class imbalance commonly found in cyberbullying datasets, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE is a widely adopted oversampling method that generates synthetic samples for the minority class by interpolating between existing data points, thereby expanding its size and reducing imbalance. In the context of cyberbullying detection, datasets often include multiple labels corresponding to specific types of harmful or benign content—such as ethnicity or race-based abuse, gender or sexual orientation-related bullying, age discrimination, religious targeting, hate speech, offensive language, and non-offensive (normal) content. These labels naturally result in an uneven distribution, with certain categories significantly underrepresented. To simplify the classification task and improve model performance, these labels were consolidated into two broad categories: **"Cyberbullying,"** encompassing all harmful or abusive content, and **"Not Bullying,"** representing neutral or non-offensive text. This transformation reframes the task as a binary classification problem. After this re-labeling, SMOTE was applied to synthetically balance the number of samples in both classes. The resulting dataset ensures equal



representation, which is critical for reducing model bias and enhancing generalization across categories.

Figure4: Before Data balancing

Figures 4 and 5 illustrate the class distribution before and after applying the balancing technique, highlighting the effectiveness of SMOTE in preparing the dataset for fair and accurate model training.

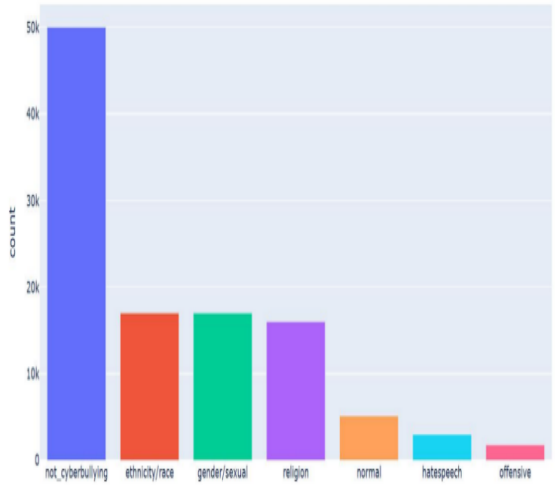


FIG5. After Data balancing

L7 Model Training

In this study, the dataset is partitioned using an 80:20 split, with 80% allocated for training and 20% reserved for testing. From the training portion, a further 80:20 split is applied to create separate training and validation sets. This ensures that the model is evaluated on previously unseen data during training, promoting better generalization and reducing overfitting. The core focus of this research is on transformer-based architectures, specifically implementing a hybrid model that integrates XLNet and RoBERTa to enhance classification performance.Both XLNet and RoBERTa are pretrained on extensive text corpora and are known for their strong contextual understanding capabilities. In the proposed hybrid framework, the models operate in parallel: input text is tokenized separately for each transformer and processed independently through their respective encoders. The resulting embeddings from both models are then concatenated and passed through a shared classification head to generate the final prediction.Training is carried out using the AdamW optimizer, along with a cosine learning rate scheduler and learning rate warmup, which helps stabilize training in the initial stages. Model performance is monitored using accuracy as the primary evaluation metric on both training and validation datasets. Once training is complete, the best-performing model is saved for deployment and future inference tasks. Figure 6 illustrates the complete pipeline, including data splitting, model architecture, and training workflow.

Figure6: Pipeline overview

B. Performance Evaluation

The hybrid XLNet–RoBERTa model demonstrates strong and reliable performance in detecting cyberbullying content. Throughout training, the model's validation accuracy improved consistently across epochs, ultimately reaching 96.76%, accompanied by a minimal validation loss of 0.0941. These results indicate excellent generalization capabilities, suggesting the model is well-suited to handle previously unseen data. When evaluated on a balanced, real-world test set comprising 10,000 samples, the system achieved an overall accuracy of 97%, as detailed in the classification report.A key strength of this hybrid architecture lies in its balanced precision and recall scores—both ranging from 96% to 98%—across cyberbullying and non-cyberbullying categories. The model consistently maintains high F1-scores, minimizing both false positives and false negatives. This level of precision and consistency makes the system particularly suitable for real-time deployment, where

both accuracy and reliability are critical.Figures 8 and 9 provide a visual summary of the model's training dynamics. Figure 8 illustrates the progression of training and validation accuracy over multiple epochs, while Figure 9 presents the corresponding training and validation loss curves, offering insight into the model's learning behavior and its ability to generalize effectively.

Figure7: Training vs. Validation Accuracy across epochs

Figure8: Training vs. Validation Loss across epochs

To assess the effectiveness of the proposed XLNet-RoBERTa hybrid model for cyberbullying detection, we evaluated its performance over three training epochs. The model demonstrated consistent improvement in both training and validation metrics, as detailed in fig9.

Epoch	**Train loss**	**Train Accuracy**	**Validation loss**	**Validation Accuracy**
1	0.2232	90.72 %	0.1396	94.71 %
2	0.1132	95.74 %	0.0995	96.35 %
3	0.0677	97.51 %	0.0941	96.76 %

Figure9: XLNet-RoBERTa Epoch-wise Performance

After completing training, the hybrid model was tested on a balanced dataset of 10,000 samples (5,000 "Cyberbullying" and 5,000 "Not Cyberbullying"). The classification report revealed excellent precision, recall, and F1-score values across both classes, as shown in Table 3.

Formulas:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

Support for class i= Number of true samples of class i

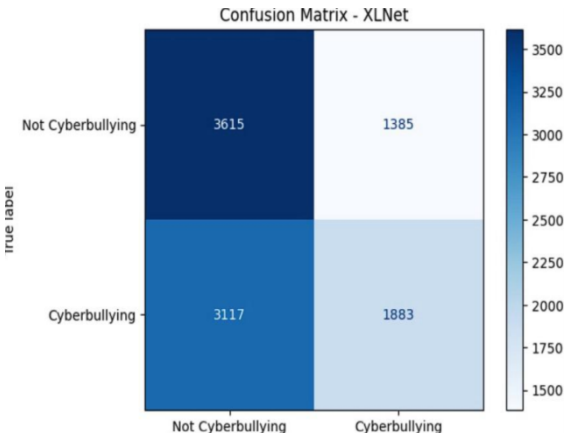
Classification report is one of the performance measurement metrics of a multi-class classification model. It indicates the accuracy, recall, F1 score, and support of the model.

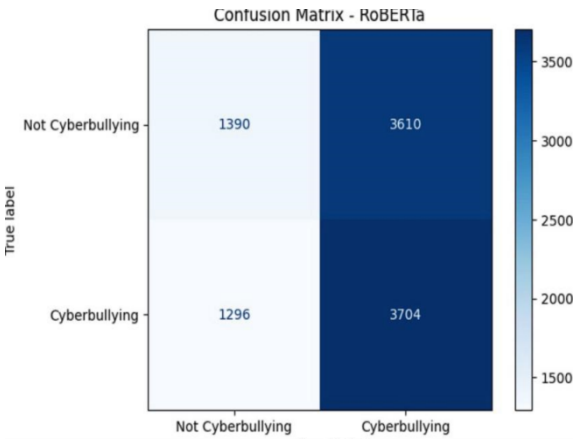
	precision	Recall	F1-Score	Accuracy
XLNet	0.5566	0.5498	0.5359	0.5498
RoBERTa	0.5120	0.5094	0.4816	0.5094
XLNet+RoBERTa	0.9677	0.9676	0.9676	0.9676

Figure :10 Classification Report on Test Data

The confusion matrix offers a clear and intuitive representation of the XLNet–RoBERTa model's classification performance across the two categories: cyberbullying and

non-cyberbullying. It reveals a high count of correct predictions for both classes, with only a small number of misclassifications. This demonstrates the model's strong ability to differentiate between harmful and non-harmful content, further validating its high precision, recall, and overall accuracy in practical scenarios. Figure 9 presents the confusion matrix, highlighting the effectiveness of the hybrid model in distinguishing between "Cyberbullying" and "Not Cyberbullying" posts. The visual clearly reflects the robustness and reliability of the system, reinforcing its suitability for real-world deployment.





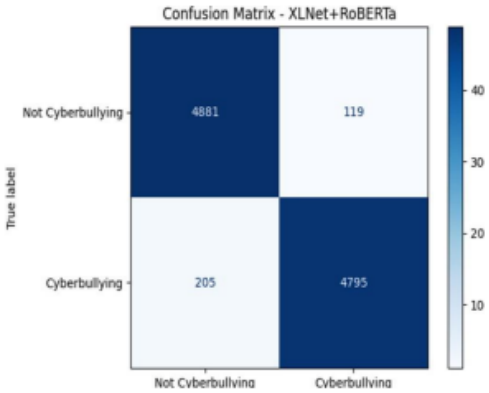


Figure11: confuison matrix

Hybrid Model Architecture

To accurately detect cyberbullying in textual data, a hybrid deep learning model was developed by integrating two advanced transformer-based architectures: XLNet and RoBERTa. XLNet is utilized for its ability to capture rich bidirectional contextual relationships through permutation-based language modeling, while RoBERTa enhances language comprehension using dynamic masking and optimized pre training strategies. In the proposed architecture, each input text is tokenized and encoded independently by both models. The resulting contextual representations—specifically the [CLS]-like tokens from each output—are extracted and concatenated to form a unified embedding. This fused representation is then passed through a neural classifier consisting of two fully connected layers, incorporating ReLU activation functions and dropout for regularization. The combination of these components enables the model to effectively leverage the complementary strengths of XLNet and RoBERTa, enhancing its ability to distinguish between cyber bullying and non-cyberbullying content with greater accuracy and robustness.

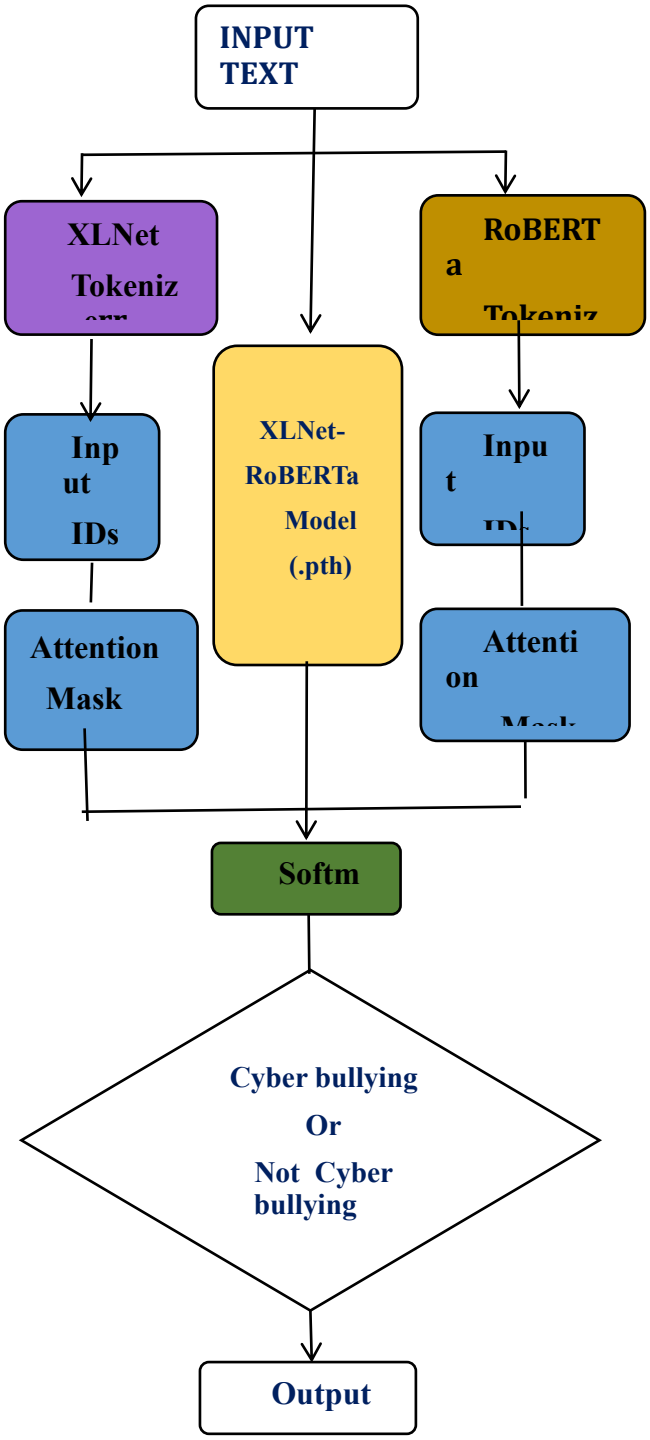


Figure12: HybridModel Architecture

Gradio-based Interactive UI with Enhanced Visual Feedback

To facilitate intuitive, real-time interaction with the cyberbullying detection model, a web-based user interface was developed using Gradio, an open-source Python framework for deploying machine learning applications. The interface was designed not only to display prediction results but also to enhance user experience through dynamic visual feedback. Upon entering a social media comment, the system invokes a custom prediction function (predict_with_color), which returns an HTML-styled output box. This box changes its background color based on the prediction result: red for "Cyberbullying" and green for "Not Cyberbullying," enabling users to instantly interpret the system's response.

The interface features a prominent header, descriptive subtext, and a modern visual aesthetic supported by a linear gradient background in light pink tones. The "Predict" button was enhanced with custom hover effects and styled using embedded CSS to improve interactivity. A footer section credits the project and the underlying platform, ensuring a professional and informative presentation.Deployment was carried out via Gradio's hosted sharing link, making the application accessible for testing on both local and remote environments. Figure 13 presents the console log output, which confirms successful execution of the Gradio app locally (http://127.0.0.1:7860) and the generation of a publicly accessible share link (https://102ecf896edb97a15a.gradio.live). This demonstrates the model's effective deployment for practical use and evaluation in real-world settings.

```
* Running on local URL:  http://127.0.0.1:7860
It looks like you are running Gradio on a hosted a Jupyter notebook. For the Gradio ap
p to work, sharing must be enabled. Automatically setting `share=True` (you can turn t
his off by setting `share=False` in `launch()` explicitly).

* Running on public URL: https://102ecf896edb97a15a.gradio.live

This share link expires in 1 week. For free permanent hosting and GPU upgrades, run `g
radio deploy` from the terminal in the working directory to deploy to Hugging Face Spa
ces (https://huggingface.co/spaces)
```

Figure13: deployment of the Gradio interface

The cyberbullying detection interface is successfully accessed on a mobile device using the Gradio public share URL(https://102ecf896edb97a15a.gradio.live). This figure demonstrates the model's accessibility across platforms, enabling users to test social media text classification for cyberbullying detection directly from smartphones through a web browser without needing to install any application.

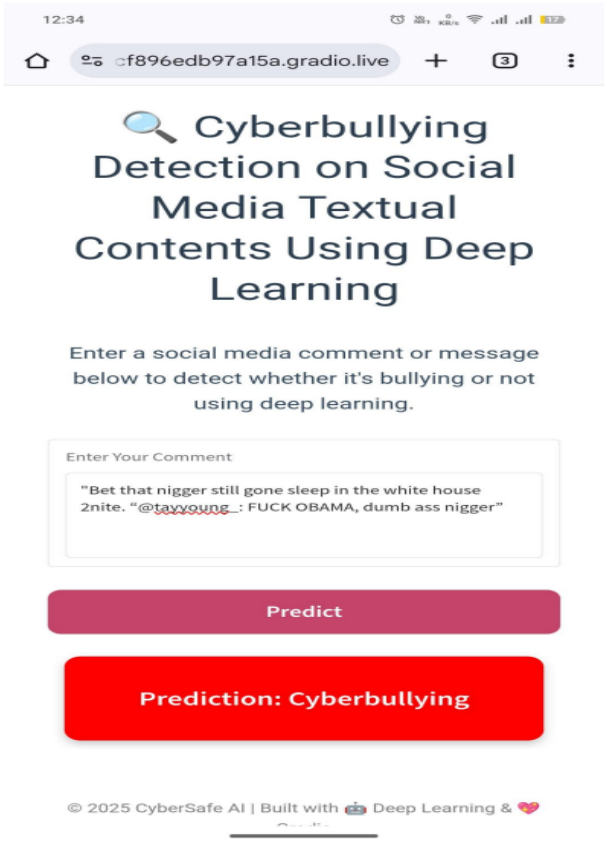


Fig 14.Cyberbullying Detection Web App Running on Mobile via Public Host

The cyberbullying detection application is running locally on a web browser via the Gradio interface, accessible at http://127.0.0.1:7860. This local instance demonstrates the developer testing phase, where the trained deep learning model classifies social media comments in real-time without requiring an internet connection or public deployment. The interface runs inside a Jupyter environment and reflects accurate backend integration with frontend visual output.

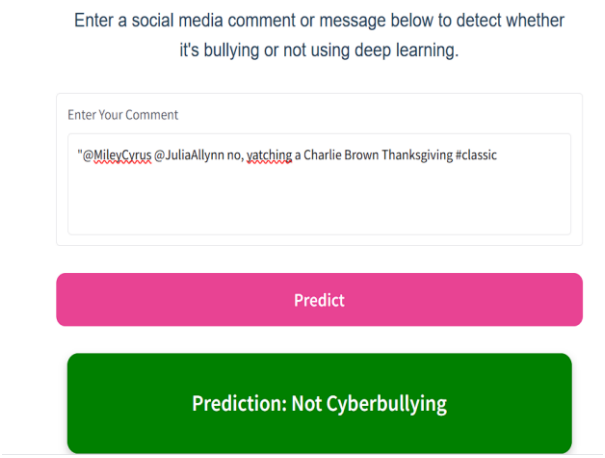


Figure 15: Cyberbullying Detection Web App Running on Localhost

Figure 16: Temporary URLs generated by Gradio during a Jupyter Notebook session. The interface is accessible via both local and public URLs, which expire once the session ends.

Fig16. After the session

V. CHALLENGES AND FUTURE WORK

While the hybrid XLNet–RoBERTa model has shown strong performance in detecting cyberbullying in social media text, several challenges remain that must be addressed to improve robustness and ensure real-world applicability. One of the primary issues is the difficulty in handling informal, slang-rich, and code-switched language, which is prevalent across social media platforms. These non-standard linguistic patterns can limit the effectiveness of even fine-tuned pre-trained models. Another key concern is the inherent imbalance and ambiguity in cyberbullying datasets. Subtle or implicit forms of abuse are often difficult to label consistently, and variations in human annotation can impact the quality of training data and model generalization.

Additionally the high computational demands and slower inference speeds of large transformer models, such as XLNet and RoBERTa, present obstacles for real-time deployment, particularly on resource-constrained devices. To address these limitations, future work will focus on enhancing model efficiency and contextual awareness. One promising direction is the integration of knowledge-enhanced pretraining or emotion-aware transformer architectures to better detect nuanced bullying cues. Further improvements will include the use of multilingual and multimodal datasets—combining text and images—to support broader generalization across different languages, user communities, and social platforms.

Finally, deploying the system as part of a real-time, cloud-based moderation tool with built-in user feedback mechanisms can help continuously refine predictions, making the model more adaptive, context-aware, and socially responsible over time.

VI. CONCLUSION

This paper presents an intelligent and effective system for detecting cyberbullying in social media text by leveraging the combined capabilities of XLNet and RoBERTa, two state-of-the-art transformer-based language models. By fine-tuning on a carefully labeled dataset and applying comprehensive preprocessing and tokenization techniques, the proposed hybrid model demonstrates enhanced accuracy and contextual understanding in identifying harmful and abusive content online. The integration of a Gradio-based user interface further improves usability, allowing real-time interaction and classification through an accessible web platform.

The results underscore the model's potential in supporting content moderation, aiding social media platforms, and promoting awareness initiatives aimed at reducing online harassment. While challenges such as ambiguity in offensive language, scalability, and deployment on low-resource environments remain, the system provides a strong foundation for future advancements. Planned improvements include enhancing computational efficiency.

expanding multilingual capabilities, and integrating the model into broader content moderation frameworks.

Overall, this work contributes meaningfully to the ongoing effort to foster safer digital environments through responsible and ethical applications of artificial intelligence in natural language processing.

VII.REFERENCES

[1] Cagirkan and Bilek conducted a study exploring the prevalence and psychological impact of cyberbullying among Turkish high school students, emphasizing the need for preventive measures in educational settings.

[2] Chi et al. examined the relationship between screen time, exposure to cyberbullying, and coping strategies among students in Hanoi, providing insight into behavioral patterns in adolescent populations.

[3] López-Martínez and colleagues proposed an automated detection system, CyberDect, which identifies bullying patterns on Twitter using advanced data mining and machine learning techniques.

[4] Kowalski and Limber analyzed the psychological, academic, and physical effects of both traditional and online bullying, offering a comparative framework for understanding the broader implications of cyberbullying.

[5] Huang compared foundational learning theories by Piaget and Vygotsky, highlighting how these educational perspectives can relate to the development of digital literacy and online behavior.

[6] Anwar et al. explored the links between workplace cyberbullying and interpersonal deviance, identifying emotional exhaustion and silence as

mediating factors in organizational environments.

[7] Kee and co-authors investigated the influence of the COVID-19 pandemic on the frequency and nature of cyberbullying on social platforms,

[8] highlighting shifts in digital behavior during global crises.

[9] Kwan et al. conducted a systematic review to map the effects of cyberbullying on children's mental health, providing a foundation for policy development and digital wellness initiatives.

[10] Garett et al. reviewed existing literature on the relationship between social media use and exposure to cyberbullying, identifying risk factors and behavioral trends.

[11] Ptaszynski and team developed a method for extracting harmful sentence structures from online conversations to improve automated cyberbullying detection models.

[12] In a follow-up study, Ptaszynski et al. presented a brute-force technique for pattern extraction that significantly enhances message classification accuracy in cyberbullying contexts.

[13] Raza and co-authors implemented supervised machine learning techniques to detect offensive content in public comments, contributing to scalable online moderation systems.

[14] Nguyen and colleagues analyzed how text reflects social and cultural behavior, providing insights into computational sociolinguistics relevant to cyberbullying detection.

[15] Cai et al. demonstrated the effectiveness of deep learning models in text classification tasks, supporting the growing role of neural networks in natural language processing applications.

[16] Tiku and Newton reported on Twitter's internal acknowledgment of its shortcomings in handling abuse on the platform, underscoring the practical need for automated moderation solutions.