iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78853

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 III Mar 2026- Available at www.ijraset.com

Cyberbullying and Hate Speech Detection Using DistilBERT-Based NLP Techniques

Pavan Dolle[1], Atharv Gadekar[2], Sunny Gend[3], Om Gojamgunde[4]
Department of Computer Engineering JSPM University, Wagholi, Pune – 412207, Maharashtra, India

Abstract: ***The rapid growth of online social platforms has increased digital communication, but it has also led to an increase in cyberbullying and hate speech. These harmful interactions can have negative effects on individuals and reduce the safety of online environments. Due to the large amount of user-generated content, manual monitoring is not feasible, making automated detection systems essential. This paper presents a system for detecting cyberbullying and hate speech using Natural Language Processing (NLP) and the DistilBERT model. The proposed approach categorizes text into four categories: non-cyberbullying, religion-based hate, gender-based hate, and racism. A dataset of approximately 100,000 samples collected from Kaggle is used for training and evaluation. The input text is preprocessed using standard NLP techniques, including text cleaning and tokenization, before being passed to the DistilBERT model for classification. The system is further integrated with a Flask-based web application, allowing users to test the model in real time and receive immediate feedback. Experimental results show that the model achieves an accuracy of 99.6% along with high precision, recall, and F1-score values. The system effectively captures the meaning of text, allowing it to outperform traditional machine learning approaches. Additionally, the implementation of an automatic restriction mechanism helps prevent repeated posting of harmful content. Overall, the proposed system provides an efficient and practical solution for automated content moderation and contributes to improving the safety of online communication platforms.***

Index Terms: ***Cyberbullying, Hate Speech Detection, Natural Language Processing, DistilBERT, Deep Learning, Text Classification***

I. INTRODUCTION

The widespread use of social media platforms has changed how people communicate and share information.While these platforms make communication easier, they have also con- tributed to the spread of cyberbullying and hate speech, posing serious challenges to online safety and user well-being. These harmful forms of communication can have negative psychological and social effects, making the development of effective detection mechanisms essential. Cyberbullying refers to the misuse of digital communication platforms such as social media, messaging services, and online forums to harass or target individuals [1]. It typically involves offensive language, threats, and discriminatory expressions based on attributes such as religion, gender, or race. Due to its ability to spread quickly and reach a large audience, cyberbullying has a more severe impact compared to traditional forms of bullying.

Content on social media often includes informal elements such as slang, abbreviations, emojis, and context-dependent expressions, which makes detection challenging for traditional machine learning techniques. Recent advancements indicate that transformer-based models provide better performance than conventional methods by effectively capturing contextual and semantic relationships [2], [3].

Many studies have explored the use of advanced deep learning techniques and transformer-based architectures for hate speech detection.Various studies highlight that models like DistilBERT and BERT-based frameworks offer efficient and accurate solutions while reducing computational complex- ity [2]. Additionally, multilingual and large language model approaches have been introduced to address challenges related to diverse languages and cultural contexts [4]. Other research emphasizes the importance of explainability, efficiency, and scalability in hate speech detection systems to improve their practical usability [5].

In this work, a system for detecting cyberbullying and hate speech is developed using the DistilBERT model for multi- class text classification.The system categorizes text into four classes: non-cyberbullying, religion-based hate speech, gender- based hate speech, and racism. The input data is preprocessed using standard NLP techniques, including cleaning and tokenization, to improve data quality before classification. Fur- thermore, a Flask-based web application is integrated into the system to enable real-time interaction and instant prediction for users.

Unlike many existing approaches that focus only on de- tection, this system also includes a restriction mechanism to limit repeated harmful behavior, helping to not only identify harmful content but also reduce its spread. Instead of relying solely on high accuracy

values, the system is evaluated using multiple performance metrics such as precision, recall, and F1- score to ensure balanced and reliable performance.In addition, the system is designed to operate in real time, allowing im- mediate analysis of user input and quick response generation. This real-time capability makes it suitable for deployment in modern social media platforms where instant moderation is required. The integration of an interactive web interface further enhances user accessibility and practical usability. Moreover, the system can be extended to support additional categories of harmful content as needed. Overall, the proposed system aims to provide an efficient, scalable, and practical solution for cyberbullying and hate speech detection, contributing to safer and more responsible online communication environments.

II. LITERATURE REVIEW

A significant amount of research has been conducted in the area of cyberbullying and hate speech detection using machine learning and natural language processing techniques. This section presents key studies and existing approaches that form the basis for the development of the proposed system.

A. Zero-Shot Learning for Hate Speech Detection

Several research efforts have explored zero-shot learning techniques using large language models for hate speech de- tection without relying on labeled datasets [6].This method helps overcome challenges such as the scarcity of annotated data and the variation of hate speech across different languages and contexts. By applying prompting strategies, these models can perform classification tasks based on their pre-trained knowledge. However, the effectiveness of this approach largely depends on the design of prompts and the choice of the underlying model.

B. BERT-Based Hate Speech Detection

Transformer-based models like BERT have significantly improved performance in NLP tasks, including text classi- fication and sentiment analysis [2]. These models capture contextual relationships between words, making them more effective for hate speech detection. However, BERT requires high computational resources, leading to the development of lightweight variants such as DistilBERT.

C. Cyberbullying in Higher Education

Studies on cyberbullying in higher education highlight its psychological and social impact on students and staff [7]. These studies emphasize the need for both technological solutions and institutional policies to reduce cyberbullying incidents and support victims.

D. Challenges in Hate Speech Detection

Detecting hate speech on social media is challenging due to informal language, slang, emojis, and context-dependent meanings [8]. Deep learning approaches have been proposed to address these issues, but challenges such as data scarcity and overfitting still remain.

E. Cyberbullying Among Adults

Research shows that cyberbullying is not limited to students but also affects professionals, including university faculty members [7]. This highlights the need for detection systems that can work across different user groups and platforms.

F. NLP-Based Hate Speech Detection

Various NLP techniques, including machine learning and deep learning, have been used for detecting hate speech [9]. These methods classify different categories such as racism, sexism, and religious discrimination, but still face challenges like sarcasm and ambiguity.

G. Security and Privacy in Online Social Networks

User-generated content in online social networks can be misused for harmful activities, including cyberbullying [10]. This emphasizes the importance of automated monitoring systems and security mechanisms to protect users.

H. Limitations of Existing Research

Although significant progress has been made, most existing systems focus mainly on detection and lack preventive mech- anisms. Many approaches perform only binary classification and do not identify specific categories of hate speech. Ad- ditionally, real-time response and user-level control are often missing.

I. Proposed System Contribution

To address these limitations, the proposed system uses a DistilBERT-based model for efficient and accurate text clas- sification. The system performs multi-class classification into four categories: non-cyberbullying, religion-based hate speech, gender-based hate speech, and racism. Furthermore, a user restriction mechanism is introduced to temporarily limit users from posting harmful content, enhancing both detection and prevention.

III. METHODOLOGY

This section describes the overall working process of the proposed Cyberbullying and Hate Speech Detection System. The system is designed using Natural Language Process- ing (NLP) techniques and the DistilBERT model to classify user-generated text into multiple categories such as non- cyberbullying, religion-based hate, gender-based hate, and racism. The methodology is divided into multiple stages including system architecture design and module-wise imple- mentation.

A. System Architecture

The proposed system follows a structured pipeline archi- tecture where input text flows through multiple stages such as data preprocessing, model prediction, and action handling. The architecture ensures efficient processing and real-time detection of harmful content.

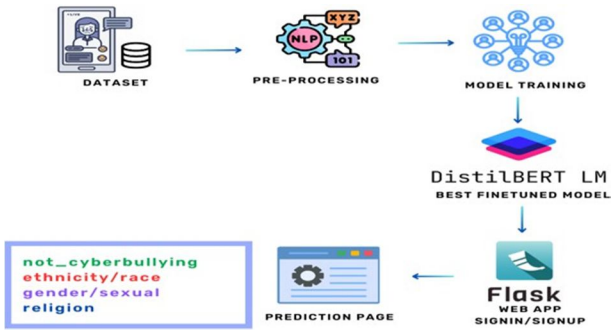


Fig. 1. Overall System Architecture of the Proposed System

The system architecture mainly consists of four major components:

- Data Collection Layer
- Data Preprocessing Layer
- Model Training and Prediction Layer
- Web Application and User Interaction Layer

The input provided by the user is first processed and cleaned before being passed to the trained DistilBERT model. Based on the prediction, the system takes appropriate action such as allowing the post or restricting the user.

B. Module Design

The system is divided into multiple modules to ensure modularity and scalability. Each module performs a specific task in the overall pipeline. This modular design enables in- dependent development and testing of each component, im- proving system reliability. It also allows easy integration of new features without affecting existing functionalities. As a result, the system becomes more flexible and adaptable to future requirements. Breaking the system into smaller parts helps in handling the overall process more effectively. Each section focuses on a specific function, making the workflow more organized. This also helps in quickly identifying and fixing issues when needed.

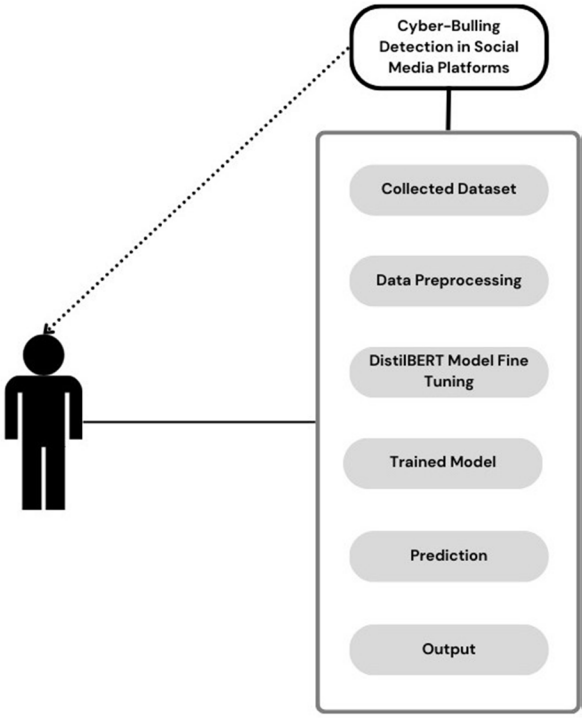


Fig. 2. Module Design Flowchart

C. Module 1: Data Collection

The first step in the system is data collection. A large dataset consisting of approximately 100,000 text samples was collected from Kaggle. The dataset includes various categories of cyberbullying and hate speech such as racism, gender bias, and religious hate.
The collected dataset plays a crucial role in training the model effectively. It contains both positive (hate speech) and negative (non-hate speech) examples, which helps the model learn meaningful patterns.

D. Module 2: Data Preprocessing

Before feeding the data into the model, preprocessing is performed to clean and normalize the text data. This step improves the performance of the model.

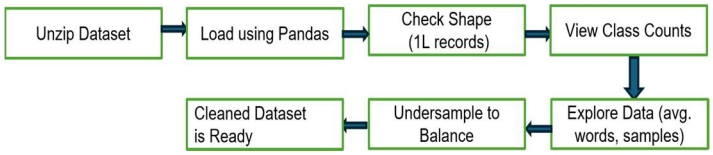


Fig. 3. Data Preprocessing Steps

The preprocessing steps include:

- Removal of special characters, URLs, and emojis
- Conversion of text to lowercase
- Tokenization of sentences
- Stopword removal
- Text normalization

These steps help to reduce noise and improve the quality of the input data for the model.

E. Module 3: Model Training using DistilBERT

The core of the system is the DistilBERT model, which is a lightweight version of BERT and provides faster performance while maintaining high accuracy.

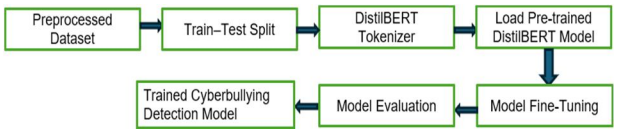


Fig. 4. DistilBERT Model Training Process

The model is trained on the preprocessed dataset to classify text into four categories:

- Non-Cyberbullying
- Religion-based Hate Speech
- Gender-based Hate Speech
- Racism

During training, the model learns contextual relationships between words using transformer architecture. Fine-tuning is performed to adapt the model to the specific task of hate speech detection.

F. Module 4: Flask Web Application (User Interface)

A Flask-based web application is developed to provide an interactive platform for users. It enables users to input text and receive predictions in real time.

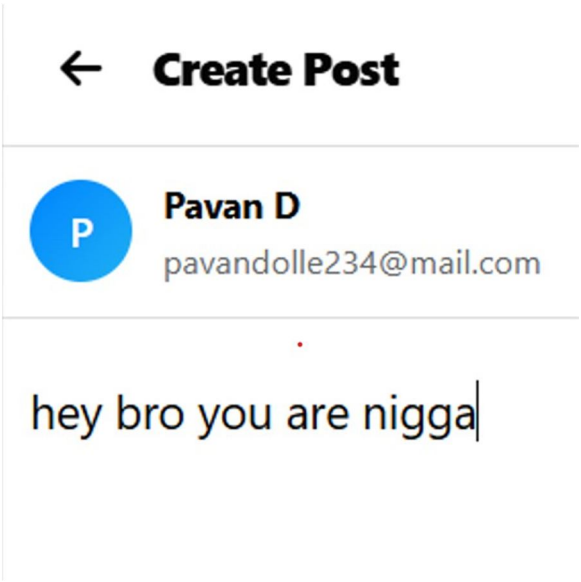


Fig. 5. User Input Interface of the Flask Application

The interface allows users to easily enter text, which is then processed by the system for analysis. It is designed to be simple and intuitive, ensuring smooth interaction for users. The system provides quick responses, allowing users to efficiently evaluate text for potential cyberbullying or hate speech.

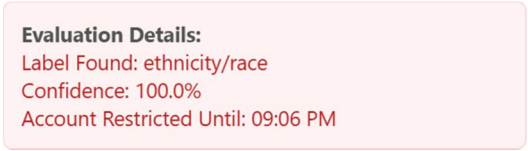


Fig. 6. Prediction Output Displayed to the User

Once the input is submitted, the system displays the classifi- cation result, such as non-cyberbullying, racism, gender-based hate, or religion-based hate.

G. Module 5: Automatic Restriction Mechanism

An important feature of the proposed system is the au- tomatic restriction mechanism, which responds immediately when harmful content is detected.

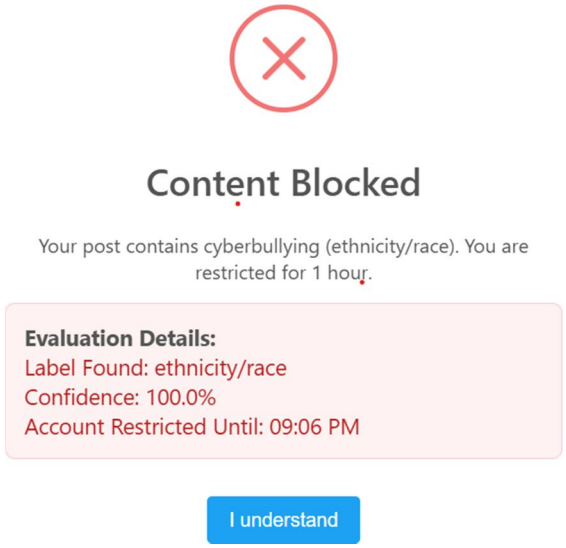


Fig. 7. Alert Message for Detected Harmful Content

When inappropriate content is identified, the system gener- ates an alert to inform the user about the violation.

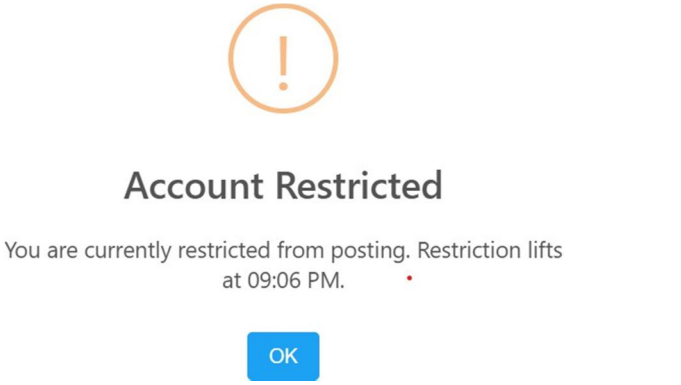


Fig. 8. User Restriction Mechanism for Detected Hate Speech

The restriction mechanism includes:

- Temporary blocking or restriction of the user
- Configurable restriction duration (e.g., 2 or 4 hours)
- Display of a warning message to notify the user

This mechanism helps in reducing repeated misuse and contributes to maintaining a safer online environment.

IV. RESULTS AND DISCUSSION

A. Confusion Matrix Analysis

The performance of the proposed model is evaluated using a confusion matrix, which provides a detailed view of classi- fication results across all classes.

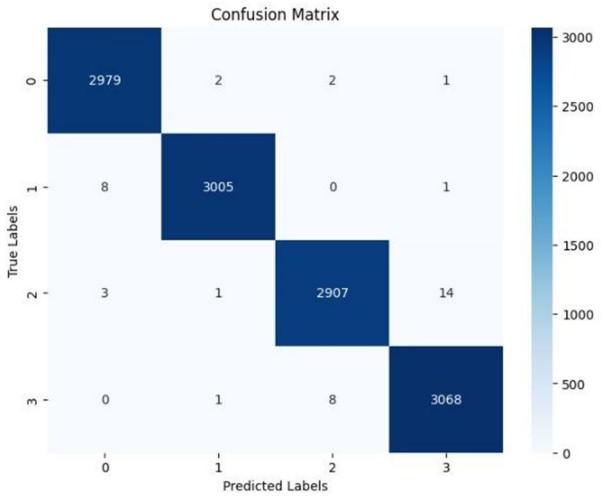


Fig. 9. Confusion Matrix of the Proposed Model

The confusion matrix shows that most predictions lie along the diagonal, indicating correct classifications [11]. Each class (0, 1, 2, 3) represents non-cyberbullying, religion-based hate, gender-based hate, and racism, respectively.

The high values on the diagonal indicate that the model correctly classifies most samples, while the off-diagonal val- ues are very small, showing minimal misclassification. This demonstrates that the model is highly accurate and can effec- tively distinguish between different categories of hate speech.

B. Performance Metrics

The model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score [12], [13].

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2984
1	1.00	1.00	1.00	3014
2	1.00	0.99	1.00	2925
3	0.99	1.00	1.00	3077
accuracy			1.00	12000
macro avg	1.00	1.00	1.00	12000
weighted avg	1.00	1.00	1.00	12000

Fig. 10. Classification Report of the Model

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

From the classification report, the model achieves near- perfect values (approximately 1.00) for precision, recall, and F1-score across all classes.

C. Training Performance

The training progress of the model is evaluated using train- ing loss, validation loss, and accuracy over multiple epochs.

[2250/2250 1:57:33, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.030500	0.023232	0.994667
2	0.000700	0.019769	0.995583
3	0.000400	0.016384	0.996583

Fig. 11. Training and Validation Performance

The results show a steady decrease in both training and validation loss, indicating effective learning. At the same time, accuracy improves with each epoch, reaching approximately 99.6% [14].

The reduction in loss values demonstrates that the model is minimizing prediction errors, while the increasing accuracy confirms improved classification performance. The small gap between training and validation loss indicates that the model is not overfitting and generalizes well to unseen data [13].

D. Discussion

The experimental results demonstrate that the proposed DistilBERT-based model achieves high performance in de- tecting cyberbullying and hate speech. The confusion matrix confirms accurate classification across all categories, while the evaluation metrics indicate balanced and reliable performance [11], [12].Compared to traditional machine learning models, the use of transformer-based architecture enables better un- derstanding of contextual relationships in text. Additionally, the integration of the model with a real-time web application enhances its practical usability.Overall, the system provides an efficient and scalable solution for automated content modera- tion and contributes to creating a safer online environment.

Another important aspect of the proposed system is its real- time implementation capability through the Flask-based web interface. Unlike traditional offline models, this system allows users to input text and receive immediate predictions, making it highly suitable for practical deployment. The integration of an alert mechanism further enhances user awareness by notifying them when harmful content is detected, thereby promoting responsible communication behavior.

Furthermore, the inclusion of an automatic restriction mech- anism significantly strengthens the system's effectiveness in real-world applications. By temporarily blocking users who attempt to post harmful content, the system not only detects but also actively prevents the spread of cyberbullying. This proactive approach distinguishes the proposed system from conventional detection-only models and contributes toward building a safer and more controlled digital environment.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper, a cyberbullying and hate speech detection system based on the DistilBERT model is proposed and imple- mented. The system is designed to classify user-generated text into multiple categories, including non-cyberbullying, religion- based hate, gender-based hate, and racism.

By leveraging the capabilities of transformer-based natural language processing, the model effectively captures contextual relationships in textual data, leading to highly accurate classification results.

The experimental evaluation demonstrates that the pro- posed model achieves high performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. The confusion matrix and training results indicate that the model generalizes well to unseen data and maintains consistent performance across different categories.

A key contribution of this work is the integration of the trained model with a Flask-based web application, enabling real-time user interaction and prediction. Furthermore, the system incorporates an automatic restriction mechanism that temporarily blocks users who attempt to post harmful or offensive content. This feature plays a significant role in re- ducing the spread of cyberbullying and promoting responsible communication.

In real-world scenarios, the proposed system can be effec- tively deployed in social media platforms, online communities, and communication systems to automatically monitor and control harmful content. Overall, the system provides an efficient, scalable, and practical solution for ensuring safer online environments using advanced NLP techniques.

B. Future Work

The proposed system can be further enhanced in several ways to improve its performance and applicability:

1) Multilingual Support: The current system is limited to a single language. Future work can focus on extending the model to support multiple languages, enabling detection of hate speech across diverse linguistic contexts.
2) Advanced Models (GPT-based): Future improvements may include the integration of more advanced language models such as GPT-based architectures (e.g., GPT-5 level models), which can provide better contextual understanding and im- prove classification accuracy.
3) Sarcasm Detection: Enhancing the model to detect sarcasm and implicit hate speech can significantly improve performance in real-world scenarios.
4) Real-time Social Media Integration: The system can be integrated with platforms like Twitter or Instagram for live monitoring of user-generated content.
5) Adaptive Restriction Mechanism: The restriction sys- tem can be improved by introducing dynamic penalties based on user behavior and repeated violations.
6) Larger Dataset Utilization: Using larger and more diverse datasets can help improve model robustness and gen- eralization.
7) Explainable AI: Future work can include explainable AI techniques to provide reasons behind model predictions.
8) Mobile Application Integration: Developing a mobile- based application can increase accessibility and usability of the system.

VI. ADVANTAGES, LIMITATIONS, APPLICATIONS AND FINAL REMARKS

A. Advantages

- High accuracy due to the use of DistilBERT model
- Real-time detection of cyberbullying and hate speech
- Automatic restriction mechanism to prevent harmful con- tent
- User-friendly web interface using Flask
- Efficient and scalable system for real-world deployment

B. Limitations

- Limited understanding of sarcasm and contextual nuances
- Performance depends on dataset quality
- Limited support for multiple languages
- Requires computational resources for training
- May misclassify ambiguous or complex text

C. Applications

- Social media platforms for content moderation
- Online gaming platforms to prevent abusive communica- tion
- Educational platforms for safe interaction
- Workplace communication systems to avoid harassment
- Comment filtering systems for websites and forums

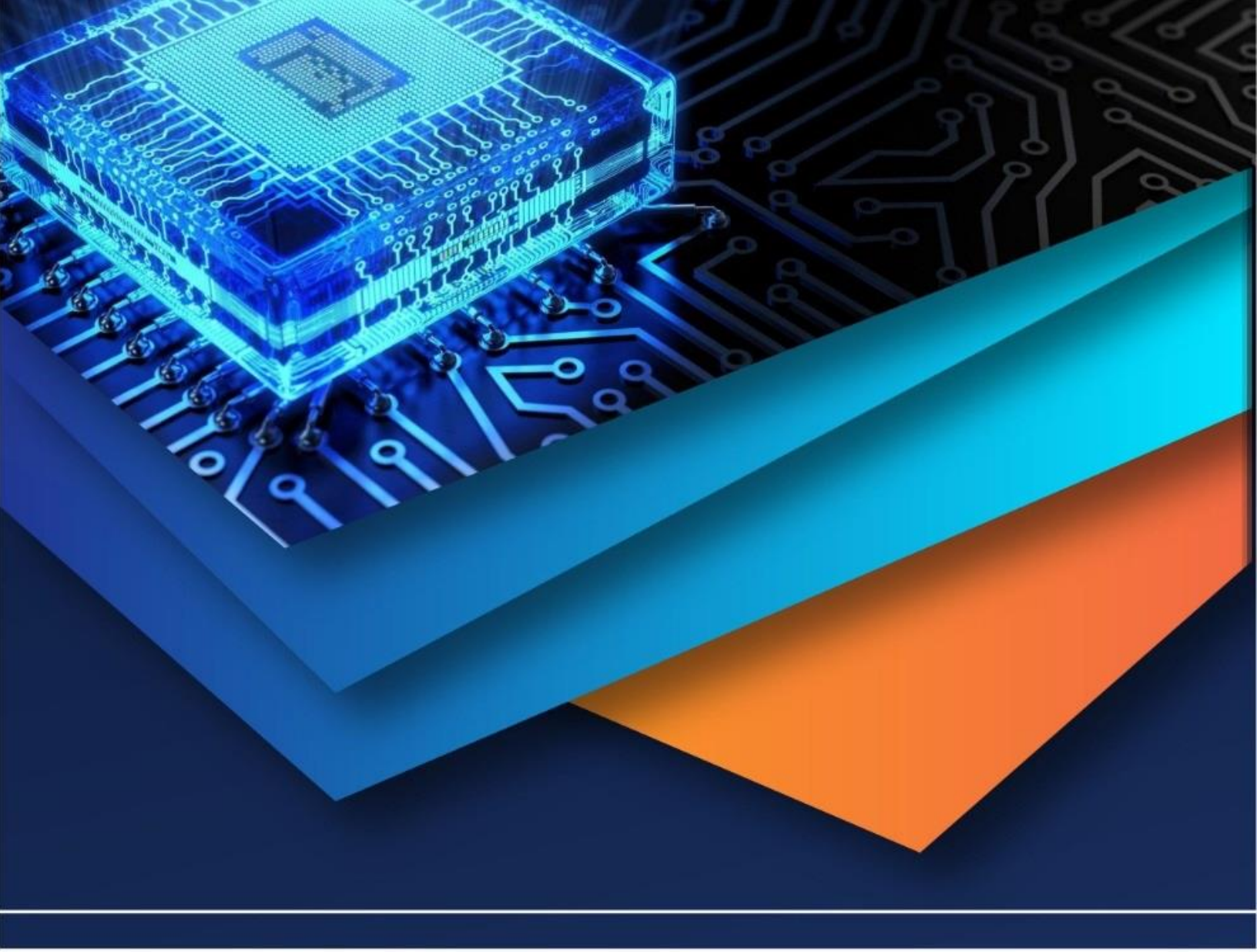
D. Final Remarks

The proposed system demonstrates the effectiveness of advanced natural language processing techniques in addressing the growing issue of cyberbullying and hate speech. With high accuracy, real-time implementation, and an automated restriction mechanism, the system offers a practical and scal- able solution for modern digital platforms. It highlights the potential of AI-driven systems in promoting safer and more responsible online communication environments.

REFERENCES

[1] S. Kowalski, S. P. Limber, and P. W. Agatston, Cyberbullying: Bullying in the Digital Age, 2nd ed. Hoboken, NJ, USA: Wiley- Blackwell, 2012.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, pp. 4171–4186.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ”DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019.
[4] A. Aluru, B. Mathew, P. Saha, and A. Mukherjee, ”Deep Learning Models for Hate Speech Detection on Social Media: A Review,” ACM Computing Surveys, vol. 54, no. 3, pp. 1–38, 2021.
[5] Z. Sokolova´, M. Grigorev, and P. Hrku´t, ”Recent Trends in Hate Speech Detection Using Natural Language Processing,” Acta Electrotechnica et Informatica, vol. 22, no. 2, pp. 54–62, 2022.
[6] F. M. Plaza-del-Arco, M. T. Mart´ın-Valdivia, L. A. Uren˜a-Lo´pez, and R. M. Crespo, ”Detecting Hate Speech Using Zero-Shot Learning with Language Models,” Expert Systems with Applications, vol. 213, 2023.
[7] A. Bussu and S.-A. Ashton, ”Cyberbullying and Cyberstalking in Higher Education: A Study of Online Harassment,” Journal of Further and Higher Education, 2023.
[8] G. Kova´cs, ”Challenges of Hate Speech Detection in Social Media,” in Proc. Int. Conf. on Recent Advances in Natural Language Processing, 2021.
[9] A. S. Parihar, ”Hate Speech Detection Using Natural Language Process- ing Techniques,” International Journal of Computer Applications, vol. 183, no. 42, pp. 1–6, 2021.
[10] A. K. Jain, ”Security and Privacy Issues in Online Social Networks: A Survey,” International Journal of Computer Science and Information Security, vol. 19, no. 2, pp. 45–52, 2021
[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ”Automated Hate Speech Detection and the Problem of Offensive Language,” in Proc. Int. AAAI Conf. on Web and Social Media (ICWSM), 2017, pp. 512–515.
[12] Z. Waseem and D. Hovy, ”Hateful Symbols or Hateful People? Predic- tive Features for Hate Speech Detection on Twitter,” in Proc. NAACL Student Research Workshop, 2016, pp. 88–93.
[13] P. Fortuna and S. Nunes, ”A Survey on Automatic Detection of Hate Speech in Text,” ACM Computing Surveys, vol. 51, no. 4, pp. 1–30, 2018.
[14] H. Zhang, M. Li, and Y. Liu, ”Deep Learning-Based Text Classification: A Comprehensive Review,” IEEE Access, vol. 8, pp. 204472–204492, 2020.
[15] Y. Kim, ”Convolutional Neural Networks for Sentence Classification,” in Proc. EMNLP, 2014, pp. 1746–1751.
[16] A. Joulin et al., ”Bag of Tricks for Efficient Text Classification,” in Proc. EACL, 2017, pp. 427–431.
[17] F. Pedregosa et al., ”Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[18] A. Vaswani et al., ”Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998– 6008.
[19] Kaggle, ”Hate Speech and Offensive Language Dataset,” [Online]. Available: https://www.kaggle.com
[20] Flask Documentation, ”Flask Web Development Framework,” [Online]. Available: https://flask.palletsprojects.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY