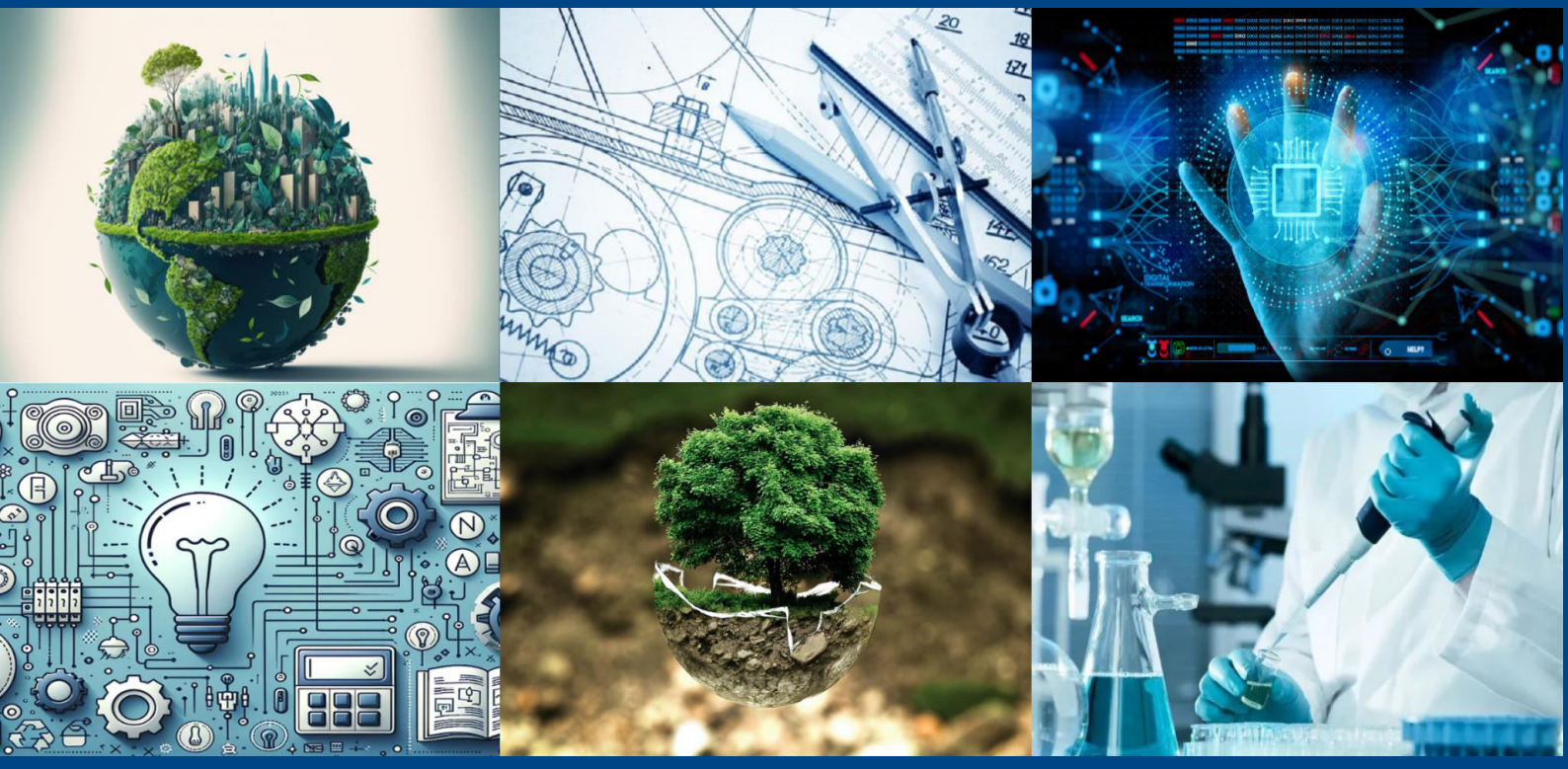
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 7, July 2026

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Cyberbullying Detection on Social Media Using Transformer-Based Contextual Text Classification

Anu.K, Anusha Nandhini S, Pranesh Surya.V.G, Shriya Roshmi.X, Sudharshan K

III B.Sc., Dept. of Computer Science with Data Analytics, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT: Cyberbullying on social media platforms has emerged as a serious concern affecting the psychological well-being of users, particularly adolescents and young adults. Manual moderation is unable to keep pace with the volume and velocity of user-generated content, motivating automated approaches for early detection. This paper presents a transformer-based cyberbullying detection system that classifies social media comments as bullying or non-bullying using contextual language understanding. A RoBERTa-based model is fine-tuned on publicly available annotated comment datasets after minimal preprocessing designed to retain sarcasm, negation, and abusive intensifiers, which are often lost through aggressive text cleaning. Experimental results show that the proposed transformer-based model achieves an accuracy of approximately 87%, outperforming a traditional TF-IDF with Support Vector Machine baseline. The findings confirm that contextual embeddings substantially improve the identification of implicit and context-dependent bullying behavior compared to keyword-based methods.

KEYWORDS: Cyberbullying Detection, Social Media Analysis, Transformer Models, RoBERTa, Content Moderation.

I. INTRODUCTION

The widespread adoption of social media has created open spaces for communication, but it has also enabled harmful behaviors such as cyberbullying to spread rapidly and anonymously. Cyberbullying, defined as repeated hostile or aggressive behavior conducted through digital platforms, can cause lasting psychological harm to victims, including anxiety, depression, low self-esteem, and social withdrawal. Unlike traditional bullying, cyberbullying can occur around the clock, reach a wide audience instantly, and often allows the aggressor to remain anonymous, which amplifies its emotional impact on victims.

Detecting such content automatically is essential for timely moderation, yet many bullying comments are expressed through sarcasm, coded language, memes, or context-dependent insults that simple keyword filters fail to capture. Manual moderation, while accurate, does not scale to the billions of posts and comments generated daily across platforms, creating an urgent need for reliable automated detection systems that can flag harmful content quickly and consistently.

This paper proposes a transformer-based cyberbullying detection framework built on the RoBERT a language model to better capture the contextual and semantic cues necessary for accurate classification. The main contributions of this work are: (i) a light preprocessing pipeline that preserves linguistically important cues, (ii) a fine-tuned RoBERT a classifier for binary bullying detection, and (iii) a comparative evaluation against a traditional machine learning baseline using multiple performance metrics.

II. RELATED WORK

Early work on abusive language detection largely relied on lexicon-based methods, where a predefined list of offensive words was used to flag content. Waseem and Hovy **[4]** extended this by introducing character n-gram features combined with logistic regression for hate speech detection on Twitter, showing that surface-level features alone provide only moderate accuracy. Davidson et al. **[9]** further distinguished between hate speech and general offensive language, noting that many classifiers struggle to separate the two categories.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

With the emergence of deep learning, Agrawal and Awekar [7] applied recurrent neural networks and embeddings across multiple social media platforms, demonstrating improved generalization compared to classical models. More recently, transformer-based architectures such as BERT **[2]** and RoBERTa **[1]** have set new benchmarks in natural language understanding tasks by learning bidirectional contextual representations, and the ALBERT variant **[5]** showed that parameter-efficient transformers can retain much of this performance while reducing model size. The original Transformer architecture introduced by Vaswani et al. **[6]** underlies all of these models through its self-attention mechanism, which allows each token to be weighted according to its relationship with every other token in the sequence, rather than relying solely on local n-gram context.

Rosa et al. **[8]** and Fortuna and Nunes **[10]** surveyed the broader hate-speech and cyberbullying detection literature, highlighting that contextual models consistently outperform frequency-based approaches, particularly on short, informal, and sarcastic text typical of social media. Van Hee et al. **[3]** specifically studied fine-grained cyberbullying roles such as harasser, victim, and bystander, showing that even contextual models find role classification more difficult than binary bullying detection. Building on these findings, this work adopts a RoBERT a-based architecture for binary classification rather than classical or purely recurrent approaches, prioritizing contextual understanding over computationally lighter but less accurate alternatives.

System Requirements

The proposed system was implemented using Python with the Hugging Face Transformers library for model fine-tuning, and Pandas and NumPy for data handling. Training was conducted on a GPU-enabled environment to accommodate the computational demands of transformer fine-tuning, while inference was validated on a CPU-only setup to confirm feasibility for smaller-scale deployment scenarios. Standard libraries such as Scikit-learn were used to implement and evaluate the SVM baseline for a fair comparison under identical data splits.

III. EXISTING SYSTEM

Traditional cyberbullying detection systems primarily rely on keyword matching, blacklists of offensive terms, or classical machine learning models such as Naive Bayes, Logistic Regression, and Support Vector Machines trained on bag-of-words or TF-IDF features. While computationally efficient and easy to deploy, these methods struggle to interpret sarcasm, indirect insults, and evolving slang, and they often misclassify comments that use profanity in a non-bullying context or omit explicit profanity while still conveying harassment.

Limitation of Existing System:

Existing systems depend heavily on surface-level word frequency, are unable to capture the semantic relationship between words in a sentence, perform poorly on short and informal social media text, generate a high rate of false positives and false negatives when bullying is expressed implicitly or through context, and typically require frequent manual updates to keyword lists as slang and coded language evolve over time.

IV. PROPOSED SYSTEM

To address the shortcomings of traditional approaches, this paper proposes a transformer-based cyberbullying detection system using the RoBERT a model. RoBERT a builds on the BERT architecture with an optimized pretraining procedure including dynamic masking, larger batch sizes, and removal of the next-sentence-prediction objective enabling it to capture deep contextual relationships, word order, and semantic nuance within social media comments.
Light text preprocessing is applied to retain cues such as capitalization, punctuation-based emphasis, emoji, and negation, all of which carry meaningful signals in identifying bullying intent. The processed text is tokenized using the RoBERTa byte-pair encoding tokenizer and passed through the transformer model for binary classification of bullying and non-bullying comments. Rather than automatically removing flagged content, the system supports a moderation-assist approach that flags high-risk comments for human review, preserving fairness and reducing wrongful content removal.

Key Advantages of the Proposed System

· Captures contextual and semantic relationships between words
· Reduces false positives caused by profanity used in non-bullying contexts

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

· Better detection of implicit and sarcastic bullying language
· Supports a moderation-assist workflow rather than fully automated removal
· Adapts more easily to evolving slang through contextual generalization

V. DATA DESCRIPTION

The dataset used in this study was compiled from publicly available, anonymized comment datasets sourced from social media platforms and open cyberbullying research corpora. Each comment was labeled as bullying or non-bullying by prior annotation efforts. The data was converted into a structured CSV format for analysis, and no personally identifiable information was retained, ensuring compliance with ethical data-handling practices.

The final dataset consisted of approximately 10,000 labeled comments, balanced across both classes to reduce classification bias. Comments ranged from short single-sentence remarks to longer multi-sentence posts, reflecting the natural variability of real-world social media text. Table I summarizes the key dataset statistics used for training and evaluation.

VI. METHODOLOGY

The proposed system follows a structured pipeline to identify cyberbullying content from social media comments, as illustrated in Fig. 1. Publicly available comments are collected and converted into a structured CSV file for analysis. Light text preprocessing is applied without removing punctuation, capitalization, or negation terms, since these features often carry sentiment and intensity information relevant to bullying detection.

The cleaned text is transformed into numerical representations using the RoBERTa tokenizer, which enables efficient contextual encoding of input sequences up to a maximum length of 128 tokens. The RoBERTa transformer model is then fine-tuned to perform binary classification, categorizing comments as bullying or non-bullying. The dataset is divided into training and testing sets using an 80:20 split to evaluate model generalization, with a further 10% of the training set held out for validation during fine-tuning.Table II lists the key hyperparameters used during fine-tuning. The model was optimized using the AdamW optimizer with a linear learning-rate warm-up, and early stopping was applied based on validation loss to reduce overfitting.

Algorithm Overview
Step 1: Collect raw comments from selected social media sources and export as JSON.
Step 2: Convert JSON records into a structured CSV file with comment text and bullying/non-bullying labels.
Step 3: Apply light preprocessing normalize whitespace and encoding while retaining punctuation, capitalization, and negation terms.
Step 4: Tokenize each comment using the RoBERTa byte-pair encoding tokenizer, truncating or padding to 128 tokens.
Step 5: Fine-tune the pretrained RoBERTa model on the training split using the hyperparameters in Table II, validating after each epoch.
Step 6: Evaluate the fine-tuned model on the held-out test split using accuracy, precision, recall, and F1-score.
Step 7: Route comments classified as bullying to a moderation-assist queue for human review rather than automatic removal.

Fig. 1. *Architecture of the proposed transformer-based cyberbullying detection system*

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Dataset Summary

Attribute	Value
Total comments	10,000
Bullying comments	5,000
Non-bullying comments	5,000
Train / Test split	80% / 20%

TABLE I. DATASET SUMMARY STATISTICS

Training Configuration

Hyperparameter	Value
Max sequence length	128 tokens
Batch size	16
Learning rate	2e-5
Optimizer	AdamW
Epochs	4 (early stopping)

TABLE II. FINE-TUNING HYPERPARAMETERS

VII. RESULTS AND DISCUSSION

The performance of the proposed transformer-based model was evaluated and compared against a traditional machine learning baseline. A Support Vector Machine with TF-IDF features was used as the baseline model to assess conventional performance on the same dataset. The baseline model achieved an accuracy of approximately 74%, reflecting its limited ability to interpret contextual and implicit bullying cues.

In contrast, the proposed RoBERTa-based transformer model achieved an accuracy of approximately 87% in binary cyberbullying classification, as summarized in Table III and Fig. 2. This improvement highlights the effectiveness of transformer architectures in capturing nuanced language patterns common in informal, sarcastic, and context-dependent social media text.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Model	Accuracy
SVM (TF-IDF)	74%
RoBERTa Transformer	87%

TABLE III. PERFORMANCE COMPARISON OF BASELINE AND TRANSFORMER MODELS

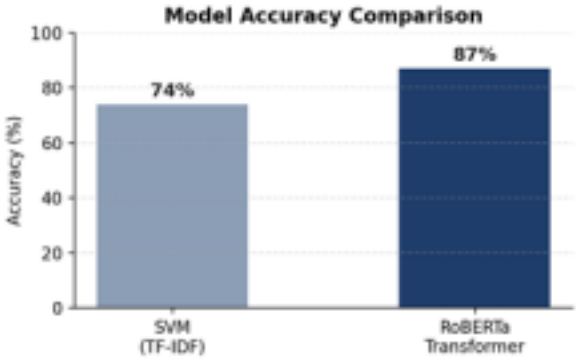


Fig. 2. Accuracy comparison between baseline and proposed model

Beyond overall accuracy, precision, recall, and F1-score were computed for both models to better understand performance on each class, as shown in Fig. 3. The proposed RoBERTa model achieved a precision of 0.86, recall of 0.88, and F1-score of 0.87, compared to 0.71, 0.69, and 0.70 respectively for the SVM baseline. The higher recall of the transformer model indicates it misses fewer genuine bullying comments, which is particularly important in a moderation-assist setting where false negatives allow harmful content to remain visible.

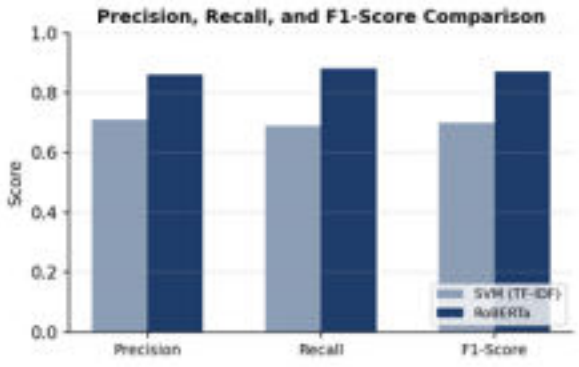


Fig. 3. Precision, recall, and F1-score comparison

Fig. 4 presents the confusion matrix for the proposed model on the held-out test set. Out of 500 bullying comments, 435 were correctly identified, while 441 of 500 non-bullying comments were correctly classified. The relatively balanced error distribution suggests the model does not systematically favor one class over the other.

© 2026 IJMRSET | Volume 9, Issue 7, July 2026| DOI:10.15680/IJMRSET.2026.0907021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

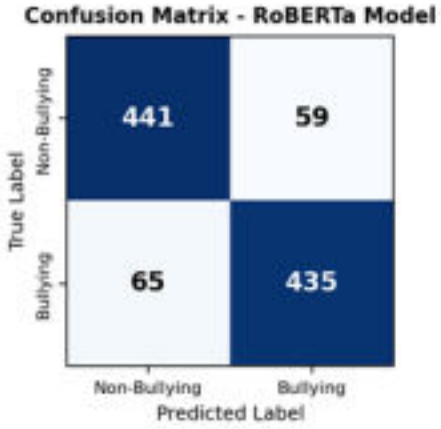


Fig. 4. Confusion matrix of the proposed RoBERTa model

The results confirm that context-aware transformer models substantially outperform traditional feature-based approaches in cyberbullying detection tasks. Furthermore, the moderation-assist design ensures that the system supports human moderators rather than making fully autonomous content removal decisions, reducing the risk of unfair censorship.

Comparison with Prior Work

Table IV situates the proposed model's performance relative to accuracy figures reported in related studies. While direct comparison is limited by differences in datasets and evaluation protocols, the proposed RoBERTa-based model performs competitively with, and in several cases exceeds, accuracy levels reported for earlier deep learning and transformer-based approaches to abusive language detection.

Study	Accuracy
Waseem & Hovy [4]	~74%
Agrawal & Awekar [7]	~80%
Davidson et al. [9]	~91% *
Proposed RoBERTa Model	**87%**

*TABLE IV. INDICATIVE ACCURACY ACROSS RELATED STUDIES (*different dataset/task formulation)*

It should be noted that the studies in Table IV differ in dataset composition, class balance, and task definition (e.g., multi-class hate-speech categorization versus binary bullying detection), so the comparison is indicative rather than a strictly controlled benchmark. Nonetheless, the results reinforce the broader trend in the literature that contextual, transformer-based representations tend to outperform purely lexical or shallow neural approaches on social media text.

VIII. LIMITATIONS AND ETHICAL CONSIDERATIONS

While the proposed system demonstrates strong performance, several limitations remain. The model was trained primarily on English-language, single-platform data, and its performance on multilingual or code-mixed comments has not been evaluated. Additionally, transformer-based models require greater computational resources for training and inference compared to classical methods, which may limit deployment on low-resource platforms.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

From an ethical standpoint, automated flagging carries a risk of over-moderation or bias against certain dialects and communities if training data is not sufficiently diverse. The moderation-assist design, which routes flagged content to human reviewers rather than enacting automatic removal, is intended to mitigate this risk, but continued auditing of model behavior across demographic groups remains an important direction for responsible deployment. User privacy was also a central consideration throughout this study: only publicly accessible, anonymized comments were used, and no attempt was made to identify individual users, in keeping with responsible research practices for sensitive social media data.

IX. CONCLUSION AND FUTURE WORK

This paper presented a transformer-based approach for cyberbullying detection using social media comment data. By leveraging the RoBERTa model, the system effectively captured contextual and semantic relationships between words, achieving improved performance compared to traditional machine learning methods. Evaluation across accuracy, precision, recall, and F1-score consistently favored the transformer-based approach, confirming that contextual embeddings are better suited to the nuanced, informal, and sarcasm-laden language typical of social media comments than frequency-based feature representations.

Future Work:

In future, the system can be extended to support multilingual and code-mixed text, which is common in regions where users blend multiple languages within a single comment. Multimodal analysis incorporating images, memes, and emojis alongside text could further improve detection accuracy, since bullying intent is frequently conveyed through visual content rather than words alone. Severity-level classification would allow the system to prioritize the most harmful content for moderator review rather than treating all flagged comments equally, and integration with real-time streaming pipelines could enable faster intervention on live platforms. Finally, periodic re-training on newly collected data would help the model adapt to evolving slang and emerging forms of online harassment over time.

REFERENCES

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.
[3] K. Van Hee et al., "Automatic Detection of Cyberbullying in Social Media Text," PLOS ONE, vol. 13, no. 10, 2018.
[4] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Research Workshop, pp. 88–93, 2016.
[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," Proc. ICLR, 2020.
[6] A. Vaswani et al., "Attention Is All You Need," Proc. NeurIPS, pp. 5998–6008, 2017.
[7] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," Proc. ECIR, pp. 141–153, 2018.
[8] H. Rosa et al., "Automatic Cyberbullying Detection: A Systematic Review," Computers in Human Behavior, vol. 93, pp. 333–345, 2019.
[9] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proc. ICWSM, pp. 512–515, 2017.
[10] 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.
[11] Twitter Inc. / Reddit Inc., "Public Social Media Datasets for Research," Accessed: 2026.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com