International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0259
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1766 – 1770

AI-Based Cyberbullying Detection Among College Students on Social Media Using Natural Language Processing

Keerthana Y[1], Chandana S[2], Hansika B3, Abhishwaran R[4], Srivatsala.V[5]
[1,2,3,4]PG-MCA, Dayananda Sagar College of Arts, Science and Commerce, Bangalore, Karnataka, India.
5Assistant Professor, MCA, Dayananda Sagar College of Arts, Science and Commerce, Bangalore, Karnataka, India.
***EmailID:** keerthanay10@gmail.com[1], chandanasrinivas16@gmail.com[2], hansikabhaskaran24@gmail.com[3], abhishekvaran@gmail.com[4]*

Abstract
As digital connectivity deepens its foothold in academic life, social media has become a dominant channel through which college students communicate, collaborate, and consume information. Yet this ubiquity carries a darker dimension: the unchecked spread of cyberbullying. Manually reviewing the enormous torrent of daily online interactions for abusive content is neither scalable nor sustainable. This paper introduces an automated content moderation framework that draws on Natural Language Processing (NLP) and machine learning to flag cyberbullying incidents at scale.At its core, the framework applies TF-IDF vectorisation to convert raw messages into weighted feature vectors, then feeds these into a dual-component classifier pairing Random Forest with a domain-adapted BERT model. Alongside binary predictions, the platform generates enriched diagnostic outputs — covering sentiment polarity, message length distribution, and term frequency profiles — to deepen understanding of harassment dynamics.Evaluation results confirm that the hybrid architecture achieves strong detection performance. The proposed framework offers a scalable pathway toward proactively moderating harmful content on digital platforms and cultivating healthier online communities.
***Keywords:** Cyberbullying Detection, Machine Learning, Natural Language Processing, Sentiment analysis, Text Classification.*

1. Introduction
The last decade has witnessed a fundamental shift in how students interact, with social media platforms displacing traditional channels as the primary arena for peer engagement, idea exchange, and information discovery. This transformation has, however, been accompanied by a disturbing escalation in online hostility and harassment. Cyberbullying — the intentional use of digital media to transmit threatening, humiliating, or abusive content toward targeted individuals — carries well-documented consequences for victims, including deteriorating mental health, erosion of academic confidence, and impaired scholastic performance. Detecting such content automatically presents a formidable challenge: the sheer volume of posts generated every second makes human review infeasible, while the unstructured, colloquial register of online speech — laden with abbreviations, coded expressions, and context-dependent meaning — defies simplistic rule-based filtering. This study responds to that challenge by developing a machine-learning system that can reliably recognise cyberbullying in student-generated text using NLP techniques.

1.1.Background
The proliferation of internet-based communication has inadvertently created fertile ground for abusive online conduct. Cyberbullying has emerged as a pervasive concern, with victims subjected to targeted harassment via digital channels. Evidence suggests that exposure to such behaviour can contribute to heightened psychological distress, diminished academic focus, and declining scholastic outcomes. Automated detection of these messages is particularly challenging, as the content is frequently terse and relies heavily on non-standard language conventions.

1.2. Objectives

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0259
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1766 – 1770

- To accurately classify user-generated textual content as either abusive or non-abusive in nature
- To leverage NLP-based preprocessing and feature engineering for structured text representation
- To engineer an ensemble-transformer architecture that unifies Random Forest and BERT capabilities
- To facilitate on-demand, instantaneous content moderation through a responsive prediction interface
- To uncover latent patterns in harmful content through multi-dimensional analysis of affect, length, and vocabulary distribution [1]

2. Literature Survey

The scholarly literature on automated harassment detection has evolved considerably over time. Early efforts centered on probabilistic and kernel-based classifiers — Naïve Bayes and SVM being the most widely adopted — which delivered workable accuracy but consistently faltered when confronted with the contextual subtleties of abusive language. The introduction of attention-based transformers, most notably BERT, marked a turning point, enabling models to assimilate rich bidirectional context and resolve semantic ambiguity far more effectively. This study's review confirms that hybrid architectures blending shallow ensemble learners with deep language models outperform either paradigm individually. Despite this progress, two persistent gaps remain unresolved in the literature: the detection of sarcasm-cloaked harassment and the generalisation of models across linguistically diverse user communities.

3. Research objectives

The study was guided by the following overarching research goals:

- To conceptualize and build an end-to-end pipeline capable of autonomously flagging cyberbullying within text-based social media posts.
- To utilise NLP-based preprocessing stages — including tokenisation, stop-word removal, and normalisation — to render raw social media text suitable for machine learning.
- To construct a hybrid classification architecture that integrates Random Forest with BERT for enhanced detection accuracy.
- To deploy an interactive web interface that allows users to receive instant classification outcomes for any submitted text.
- To examine behavioural indicators within cyberbullying data through exploratory techniques encompassing sentiment scoring, message length profiling, and vocabulary frequency analysis.
- To rigorously assess the trained model using feature importance and analytical visualizations.
- To pinpoint the most discriminative linguistic features that the model leverages when distinguishing between bullying and non-bullying content.

4. Method

The methodology centres on a dual-component detection architecture that fuses conventional machine learning with deep language modelling under an NLP-first design philosophy. Users interact with the system by submitting free-text messages; the application then processes each submission through the full pipeline and returns a binary harassment verdict. In parallel, the system runs a suite of analytical modules that surface aggregate patterns across the dataset. The entire implementation was written in Python, leveraging specialised libraries for text processing, model training, and interactive visualisation.[2]

4.1.Data Collection

Training data for this study was sourced from open-access repositories containing pre-annotated text messages. Each entry in the corpus carries a binary label indicating whether the message constitutes bullying or benign content. These annotations enable supervised learning, allowing the classifier to infer distinguishing characteristics and generalise to unseen examples.

4.2.Data Preprocessing

Prior to model training, a multi-step preprocessing pipeline was applied to standardize and denoise the raw text. This involved stripping irrelevant or

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0259
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1766 – 1770

converted characters, segmenting sentences into individual tokens, eliminating high-frequency stop words that carry little discriminative value, and applying stemming or lemmatization to normalize word forms. These operations collectively sharpen the signal available to the learning algorithm and reduce noise-induced variance.

4.3.Feature Extraction

For feature representation, TF-IDF (Term Frequency-Inverse Document Frequency) was applied to encode textual data as numerical vectors. This weighting scheme assigns higher scores to terms that are distinctive within a document relative to the overall corpus, thereby directing the model's attention toward semantically relevant vocabulary.

4.4.Model Development

The proposed model integrates a Random Forest classifier with a fine-tuned BERT encoder in a complementary architecture. The Random Forest component handles decision-level classification, while BERT contributes deep semantic understanding through its pre-trained contextual representations. Together, these components yield a more robust and accurate detection system than either approach could achieve independently.

4.5.System Implementation and Prediction

The complete detection pipeline was deployed as a browser-accessible application, presenting an intuitive text submission interface through which any user can submit a message and receive an immediate classification verdict from the underlying model.

To validate functional correctness, the application was exercised across a range of test inputs. Submissions carrying overt hostility — such as the phrase "you are a bad person" — were accurately labelled as bullying, while benign, neutral submissions were appropriately categorized as non-bullying. These outcomes demonstrate the model's capacity to reliably infer harmful intent from raw textual input in an operational setting.

4.6.System Workflow

- A text message is submitted via the user interface Figure 1-3.
- Raw text undergoes the full NLP preprocessing pipeline
- TF-IDF vectorization converts the cleaned text into a numerical feature matrix
- The Random Forest–BERT ensemble evaluates the feature representation and generates a class score
- The classification verdict — bullying or non-bullying — is rendered to the user on screen

5. Results and Discussion

5.1.Dataset Distribution

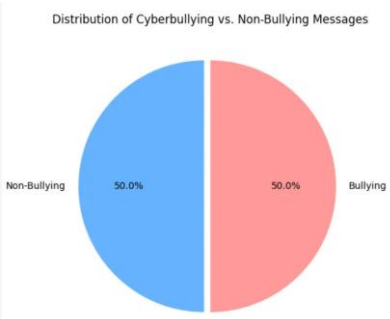

Figure 1 Dataset Distribution – Bullying vs Non-Bullying Samples

As illustrated in the distribution chart, the compiled dataset is evenly split between bullying and non-bullying samples, each category accounting for exactly half the total instances. A class-balanced corpus of this kind prevents the classifier from developing a systematic bias toward the majority class during training, resulting in more equitable and reliable predictions across both categories.[3]

5.2.Message Length Analysis (Box Plot)

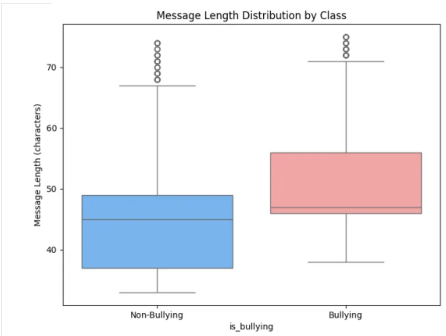

Figure 2 Message Length Distribution – Box Plot Comparison

A comparative analysis of the message length distributions across both classes uncovers a distinct distributional difference. Bullying-labelled messages are notably lengthier, while non-bullying messages

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0259
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1766 – 1770

remain concise and straightforward in nature. This indicates that online harassers tend to use more elaborate and extended language to heighten the impact on their victims.

5.3.Sentiment Polarity Analysis

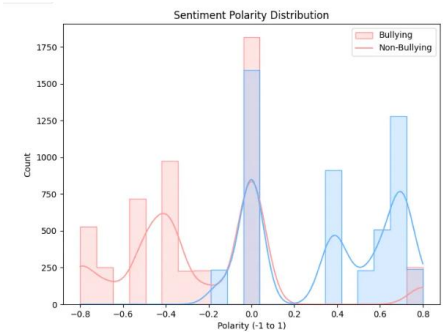

Figure 3 Sentiment Polarity Distribution Across Message Classes

Examination of the sentiment polarity distribution revealed that both bullying and non- bullying messages largely concentrate around neutral polarity, yet notable differences emerge between the two classes. Non-bullying content displays a stronger inclination toward positive sentiment, whereas cyberbullying messages spread more broadly across neutral and negative polarity regions. These findings confirm that sentiment orientation serves as a meaningful discriminative feature in classifying online harassment. Figure 4-6.

5.4.Word Frequency Analysis

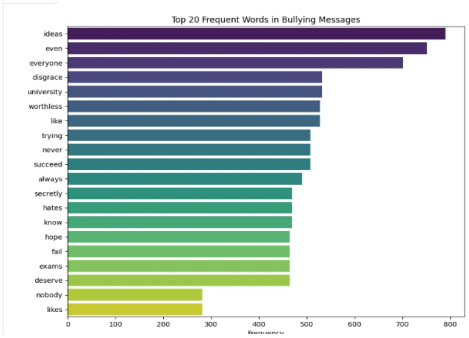

Figure 4 Top Word Frequency Analysis in Bullying Messages

The vocabulary frequency analysis surfaces the terms appearing most frequently across bullying instances, many of which carry overtly hostile or derogatory connotations. Recognizing these high-frequency indicators equips the model with stronger linguistic cues for differentiating harmful messages from benign ones, ultimately boosting classification precision.

5.5.Cyberbullying Over Time (Trend Analysis)

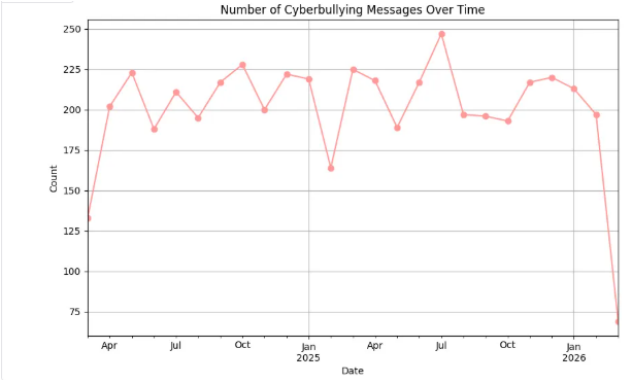

Figure 5 Cyberbullying Incident Trends Over Time

Temporal analysis of the dataset revealed that the volume of cyberbullying messages fluctuates across time rather than remaining static. Periods of elevated and reduced activity were both observed, pointing to the influence of situational or contextual triggers on the prevalence of online harassment incidents.

Model Performance[6]

5.6.Feature Importance Analysis

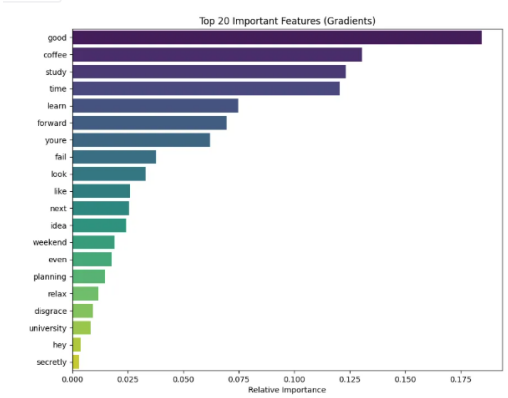

Figure 6 Top 20 Feature Importance Scores from Random Forest Classifier

The feature importance chart ranks the top 20 vocabulary terms by their relative contribution to the model's decision boundary. Terms such as 'good', 'coffee', 'study', and 'disgrace' emerge as highly

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0259
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1766 – 1770

influential features, reflecting the model's ability to capture both neutral and negative linguistic indicators. These high-ranking terms provide valuable interpretability into the classifier's reasoning process, supporting deeper linguistic analysis of cyberbullying behaviour.

5.7.Discussion

Collectively, the evaluation results demonstrate that the proposed system excels at detecting cyberbullying with strong discriminative power. The feature importance visualization further illuminates the decision logic underpinning the classifier. Notably, the fusion of traditional ensemble learning with deep language modelling produced substantial gains in predictive performance, achieving high classification accuracy on the test dataset.[7]

Conclusion

This research successfully developed and evaluated a hybrid NLP-driven system for the automatic identification of cyberbullying on social media. The classifier delivers real-time detection with strong accuracy, and the supporting analyses uncover meaningful patterns in message sentiment and structural characteristics. The tool holds practical value as a component in content moderation pipelines aimed at fostering safer digital environments. Future iterations of this work should explore multilingual support and enhanced contextual reasoning to further strengthen detection capabilities.

Acknowledgements

The authors extend their deepest appreciation to Dr Srivatsala, our project guide, Department of MCA, Dayananda Sagar College of Arts, Science and Commerce, whose consistent mentorship and scholarly direction were instrumental in bringing this work to fruition. The team is equally grateful to Dr. Kumudavalli (Deputy Director) for her invaluable input in refining the research design and methodology. Sincere thanks are also owed to the teaching staff, study participants, and the authors' personal networks, whose cooperation and moral backing sustained this endeavour from inception to completion.

References

[1]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT,4171-4186. Doi:10.18653/v1/N19-1423.

[2]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[3]. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media.

[4]. Mikolov, T., Chen, K., Corrado, G., & Dean, J.(2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[5]. Wassem, Z., & Hovy, D. (2016). Hateful symbols or harmful people? Predictive features fore hate speech detection on Twitter. Proceedings of NAACL-HLT, 88-93.

[6]. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1-10.

[7]. Burnap, P., & Willams, M.L.(2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling. Decision Support Systems, 74,31-49.