ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

Detecting Cyberbullying through Sentiment Ratings of Text and Emojis on Social Media Platforms

Dhruvi Dineshbhai Patel[1]; Vishnupant Potdar[2]; Dr. Nagnath Biradar[3]

Professor[2]

[1,2,3]Symbiosis Skills and Professional University

Publication Date: 2025/07/16

Abstract: With the growing use of digital platforms, online bullying has become a serious and widespread issue that significantly affects users' mental and emotional well-being. Public figures such as influencers and celebrities face even greater vulnerability due to their high visibility and constant exposure on online networks, especially after the sharp rise in social media usage following the pandemic. To tackle this challenge, our work adopts a well-rounded strategy that examines both the text and the expressive cues conveyed by emojis to detect cyberbullying. We utilize an array of machine learning and deep learning models—namely, Support Vector Classifier, Logistic Regression, Random Forest, XGBoost, LSTM, Bi-LSTM, GRU, and Bi-GRU to classify comments as bullying or non-bullying. Furthermore, we introduce a severity-based scoring system that rates offensive text on a scale of 1 to 5. When a message crosses a predefined severity threshold—determined by the safety standards of each platform—an automated recommendation to block the user is triggered. This approach not only enables precise identification of harmful content but also provides a proactive mechanism to promote safer online interactions.

Keywords: *Online Bullying Detection, Swear Words Dataset, TF-IDF With Random Forest, Severity Score.*

How to Cite: Dhruvi Dineshbhai Pate; Vishnupant Potdar; Dr. Nagnath Biradar (2025). Detecting Cyberbullying through Sentiment Ratings of Text and Emojis on Social Media Platforms. *International Journal of Innovative Science and Research Technology*, 10(7), 800-808. https://doi.org/10.38124/ijisrt/25jul233

I. INTRODUCTION

In today's digital-first world, the issue of online bullying—defined as the use of digital communication to threaten, insult, or harass others—has grown rapidly, leading to serious emotional and psychological effects for those affected. Individuals in the public eye, particularly celebrities and influencers, are especially prone to becoming targets due to their constant online visibility. The post-COVID era has seen a notable surge in the usage of social media platforms, which in turn has amplified both the reach and frequency of harmful online behavior.

In response to this ongoing problem, the current research proposes an all-encompassing solution for identifying and evaluating online bullying behavior across textual conversations on digital platforms. A variety of machine learning (ML) and deep learning (DL) techniques are explored and compared in this study, including Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), and XGBoost (XGB), along with neural network-based models such as LSTM, Bi-LSTM, GRU, and Bi-GRU. These models are specifically selected to process not just the textual content but also the emotional and contextual signals embedded in emojis—making the detection process more precise and human-like.

A standout aspect of this study is the incorporation of a rating mechanism that evaluates the severity of offensive content on a scale from 1 to 5. This severity score aids in making platform- specific decisions. If a message exceeds the accepted threshold value, the system automatically suggests blocking the user as a precautionary action to limit further harm. The importance of addressing online bullying lies in both its widespread nature and its damaging impact on mental health. Through the integration of accurate classification techniques, severity scoring, and actionable user recommendations, this research provides a strong framework to help social media platforms offer safer and more respectful digital spaces—especially for vulnerable individuals in the public domain.

This paper proceeds with a review of existing literature, followed by a detailed breakdown of the approach, implementation methods, experimental findings, and final conclusions. The methodology includes the use of various ML and DL models, different text encoding techniques, and well-known performance evaluation metrics to measure the

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

system's effectiveness.

II. LITERATURE REVIEW

The task of detecting online bullying has led to the development of various approaches and models, reflecting the complexity and evolving nature of harmful online interactions. This section presents a concise overview of prior studies and their significant contributions, helping to establish a foundation for understanding current advancements in the field.

Gomez et al. [12] investigated hate speech detection in multimodal content by applying different fusion techniques and alignment strategies using CNNs and RNNs. Their research incorporated textual and visual data, using methods such as Compact Bilinear Pooling and LSTM models to analyze hate speech. The study also highlighted key challenges, such as inconsistencies in dataset availability and evaluation procedures, particularly in handling multimodal social media data.

Reynolds et al. [13] focused on using language-based features to identify cyberbullying on social platforms. Utilizing machine learning tools from the Weka toolkit, their system achieved an accuracy of 78.5%. Their work emphasized the growing urgency of developing reliable detection techniques.

Mitchell et al. [14] provided a comprehensive survey addressing cyber-aggression, cyberbullying, and cyber-grooming. They employed NLP methods, traditional supervised ML, and DL models to detect subtle forms of aggression. The study advocated for real-time solutions and ethical safeguards in automated detection systems.

Salminen et al. [15] reviewed hate speech classification approaches and proposed combining datasets from multiple platforms to build stronger classifiers. Their work emphasized the importance of building systems that are effective across different social environments.

Mishra et al. [16] evaluated NLP and deep learning methods for detecting online abuse. Their study showed that transfer learning techniques improved performance, especially in detecting complex or nuanced language. Similarly, Mehendale et al. [17] explored abuse detection in Hindi- English (Hinglish) social media posts, addressing a gap in multilingual detection systems.

Neelakandan et al. [18] explored DL techniques like BCO-FSS and SSA-DBN models for cyberbullying classification, demonstrating promising results when compared with existing approaches. Their model selection focused on improving feature extraction and classification accuracy.

Talpur and O'Sullivan [19] developed a machine learning system that evaluates cyberbullying severity on Twitter. Yuvaraj et al. [20] contributed a multi-feature classification approach using deep decision tree models to process tweets and reduce overfitting.

Chia et al. [21] examined sarcasm and irony in online communication using feature engineering and ML. Their analysis considered multiple classification methods and emphasized the influence of tone and context in cyberbullying detection.

Van Hee et al. [22] focused on automated bullying detection using models such as SVM, RF, and Naive Bayes. Their promising results illustrated the value of classical algorithms when paired with well-structured data.

Perera et al. [23] proposed a detection method using Twitter data from the Internet Archive and validated their system using standard performance metrics like precision, recall, and F1-score.

Yin et al. [24] combined sentiment, content, and contextual features across platforms like Myspace, Slashdot, and Kongregate to build a harassment detection system using a linear SVM. They found that TF-IDF outperformed n-gram and foul language detection in certain cases.

Sood et al. [25] developed profanity filters using a swear word dictionary, edit distance corrections, and an SVM-based classification system. By merging the results of different detection methods, their system achieved improved accuracy in identifying offensive content.

Squicciarini et al. [26] used behavioral patterns, social interactions, and decision tree classifiers to identify bullies, introducing a rule-based logic to trace repeated offenses. Chavan and Shylaja

[27] combined skip-grams and machine learning classifiers (SVM and Logistic Regression) on Kaggle datasets to achieve improved accuracy.

In contrast to these existing systems, our study integrates a rating-based decision mechanism that can either issue a warning or suggest account restrictions depending on the platform's preferences. Additionally, our system filters out sarcasm and focuses on categorizing the core themes of bullying content for better moderation outcomes.

III. PROPOSED APPROACH

The detection pipeline begins with processing user-generated input text, which is then analyzed using a range of machine learning and deep learning algorithms such as Logistic Regression, Support Vector Classifier, Random Forest, XGBoost, LSTM, Bi-LSTM, GRU, and Bi-GRU. These models are used to classify whether the input contains cyberbullying or not, based on both language patterns and contextual cues. A visual outline of this entire process is illustrated in Fig. 1.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

Once a piece of text is flagged as containing bullying content, it proceeds to an evaluation phase where a severity score is calculated. This score is derived from an algorithm that assesses the text for elements such as intensity, use of swear words, and contextual implications. Each instance is then rated on a scale from 1 to 5, indicating how offensive it is.

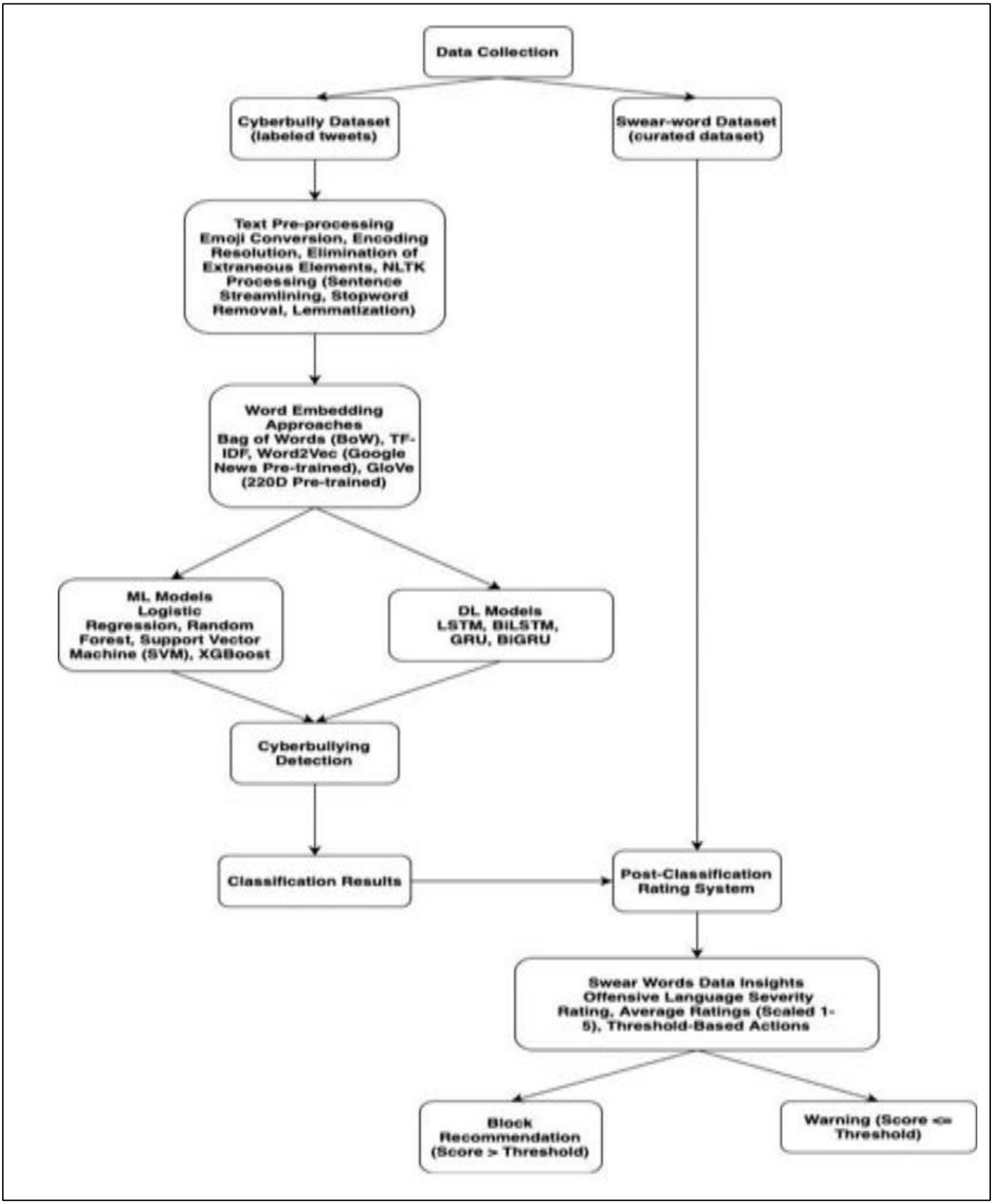

Fig 1 Flowcahart of Text Analysis for Cyberbullying

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

A decision mechanism is then applied. If the severity score surpasses a certain threshold—usually set at level 3—the system triggers an action recommending the platform to block the user as a preventive measure. On the other hand, if the score is 3 or lower, a warning message is generated instead, giving the user a chance to reflect on their behavior without immediate punitive action.

This layered approach balances automatic detection with nuanced moderation, allowing platforms to respond appropriately based on the level of severity and their specific policy guidelines.

➢ *Dataset*

The increasing frequency of cyberbullying on digital platforms has prompted researchers to create diverse datasets for effective detection. This study utilizes two primary datasets designed for identifying hate speech and offensive content.

The first dataset, introduced by Waseem and Hovy [28], consists of roughly 17,000 Tweet IDs tagged with labels addressing racism and sexism. After retrieval using the Twitter API, around 5,900 tweets could not be fetched due to deleted accounts or removed posts, resulting in a smaller usable dataset.

The second dataset, developed by Davidson et al. [29], includes approximately 25,000 annotated tweets. These tweets were categorized into three groups: hate speech, offensive language, and neutral content, based on crowd-sourced evaluations. For our study, the two datasets were merged and relabeled for simplicity: tweets marked as hate speech or offensive language were grouped under the "bully" class (1), while neutral tweets were assigned to the "non-bully" class (0). The final combined dataset includes 18,888 bullying-labeled entries and 16,081 non-bullying ones.

In addition to these, we introduced a supplementary "Swear Words Data" set, which contains 428 commonly used offensive terms found on social media. To ensure a fair and inclusive representation, this list was rated manually by individuals of different gender identities. Each term received a numerical rating based on perceived severity, helping to fine-tune the offensive scoring mechanism in our classification system. This extra layer of granularity enhances the model's sensitivity to varied expressions of aggression and supports a more robust detection framework.

➢ *Text Preprocessing*

A thorough text cleaning process was applied to prepare the data for analysis and classification. This included several important steps to standardize and simplify the input content.

To begin with, emojis present in the text were converted into their textual meanings to ensure the emotional tone embedded in these symbols could be properly interpreted by the models. Encoding inconsistencies were resolved, and all unnecessary characters—such as HTML tags, URLs, and Twitter mentions—were stripped from the text.

Next, the Natural Language Toolkit (NLTK) was used for language normalization. Common stopwords that do not add meaningful context were removed, and words were converted to their root forms through lemmatization. This step helped reduce redundancy and improve the consistency of word representations.

These preprocessing techniques ensured that the input data was uniform, cleaned, and semantically relevant, making it more suitable for embedding and classification by both ML and DL models.

- *Word Embedding Approaches*

To convert textual data into a numerical format suitable for machine learning and deep learning models, several word-embedding techniques were employed. These methods help capture the frequency, importance, and contextual relationships of words within the dataset.

- *Bag of Words (BOW):*

This basic representation technique was used to encode text based on how often each word appeared. The BoW method counts the frequency of each term within a document, creating a vectorized form of text. It serves as a solid baseline for classification tasks.

$$(d, V) = [B1, B2, \dots, BN] \quad (1)$$

Here, B(d,V) represents the document vector based on word frequencies across the predefined vocabulary V.

Term Frequency–Inverse Document Frequency (TF-IDF):

TF-IDF improves upon BoW by assigning more importance to words that appear frequently in a document but rarely across the full corpus. This allows the model to emphasize distinguishing terms over common ones.

$$TF = (\text{umber of occurrences of word "X"}) / \text{Total number of words in a}$$

$$IDF = \log(\text{Number of Documents present in a Corpus0} / \text{Number of word "X" has appeared}$$

TF-IDF=TF×IDF

- *Word2Vec (Pre-trained on Google News):*

We utilized the pre-trained Word2Vec model from Google News to understand the contextual and semantic connections between words. These vectors are based on a large news corpus and help our model understand subtle contextual meanings, improving classification performance.

- *GLOVE (220D Pre-trained):*

We also integrated the GloVe model with 220-dimensional vectors. Unlike Word2Vec, GloVe focuses on global co-occurrence statistics. This complementary embedding adds another layer of depth to word understanding, particularly for less frequent terms.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

Together, these techniques—BoW, TF-IDF, Word2Vec, and GloVe—formed a diverse embedding toolkit, enabling our models to interpret both simple frequency-based patterns and deeper semantic structures in the text.

- *Machine Learning and Deep Learning in Cyberbully Detection*

To address the complexity of identifying cyberbullying behavior, our study applies a wide range of both machine learning (ML) and deep learning (DL) models. Each of these approaches contributes unique strengths in understanding patterns and context within user-generated content.

On the ML side, we experimented with widely used algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. These models are known for their effectiveness in handling structured data and for providing interpretable results. Their capacity to detect relationships between features makes them suitable for text classification when combined with traditional embedding techniques like TF-IDF and BOW.

In parallel, we implemented deep learning models designed to handle the sequential and contextual nature of language. This includes Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), and Bidirectional GRU (Bi-GRU). These architectures are particularly adept at capturing dependencies within sentences, enabling the model to consider both past and future context in a piece of text.

The deep learning models were trained using advanced embeddings such as Word2Vec and GloVe, allowing them to learn semantic relationships more effectively. Their ability to process word sequences holistically makes them well-suited for nuanced detection tasks like cyberbullying, where tone, intent, and emotional cues play a significant role.

By combining traditional ML models with advanced DL architectures, our framework leverages the best of both paradigms, resulting in improved accuracy and a more comprehensive understanding of bullying behavior in online conversations.

- *Post Classification Rating System With Swear Words*

Following the initial classification of a text as cyberbullying, an additional assessment layer is introduced to evaluate the severity of the offense. This step relies on the curated "Swear Words Data" set, which includes commonly used abusive terms extracted from social media platforms. These terms were manually rated by individuals of diverse gender backgrounds, with each word assigned a numeric severity score based on perceived offensiveness.

Using this dataset, each bullying-labeled text is given a severity score by calculating an average based on the presence and weight of the swear words it contains. The final score ranges between 1 and 5, representing the intensity of the offensive language.

A predefined threshold, customizable based on the safety policies of different social media platforms, determines the next action. If a comment's rating surpasses this limit, the system recommends blocking the user. In cases where the score is below the threshold, a warning is issued instead. This flexible system ensures the platform's response aligns with the severity of the content, allowing for both strict intervention and milder corrections depending on context.

By incorporating swear word severity into the post-classification phase, the model not only identifies harmful content but also quantifies its potential impact. This two-step process strengthens the overall reliability and responsiveness of the cyberbullying detection framework.

IV. EMPIRICAL FINDINGS:

To evaluate the effectiveness of the proposed cyberbullying detection system, we conducted a detailed empirical analysis using a range of machine learning (ML) and deep learning (DL) models. These models were tested with various word embedding techniques to observe how text representation impacts classification performance.

In the ML experiments, we applied Logistic Regression, Support Vector Machine, Random Forest, and XGBoost, combined with different embedding strategies such as TF-IDF, Bag of Words (BoW), and Word2Vec. Each model's accuracy was calculated to measure how well it distinguished between bullying and non-bullying text. The results are summarized in Table 1.

Table 1 Accuracy Scores for Machine Learning Models

Model	TF- IDF	BoW	Word2Ve c
Logistic Regression	89.80%	90.46%	83.89%
Support Vector Machine	90.41%	84.08%	90.07%
Random Forest	93.85%	93.76%	90.51%
XGBoost	90.17%	90.00%	90.98%

As shown above, Random Forest consistently achieved the highest accuracy across all embedding types, highlighting its robustness in handling diverse feature sets.In the deep learning segment, we analyzed the performance of LSTM, Bidirectional LSTM (Bi-LSTM), GRU, and Bidirectional GRU (Bi-GRU) using GloVe and Word2Vec embeddings. The accuracy of each combination is presented in Table 2.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

Table 2 Accuracy Scores for Deep Learning Models

Model	GloVe	Word2Ve c
LSTM	89.76%	89.36%
BI- LSTM	90.10%	90.21%
GRU	89.94%	90.55%
BI-GRU	90.20%	90.03%

The results demonstrate strong performance across all DL models, with Bi-GRU and GRU slightly outperforming others when paired with Word2Vec embeddings.

Figure 2 displays the confusion matrices of top-performing models under each embedding configuration. In particular, the Decision Tree model for ML, along with LSTM and Bi-GRU for DL, showed high precision and recall.

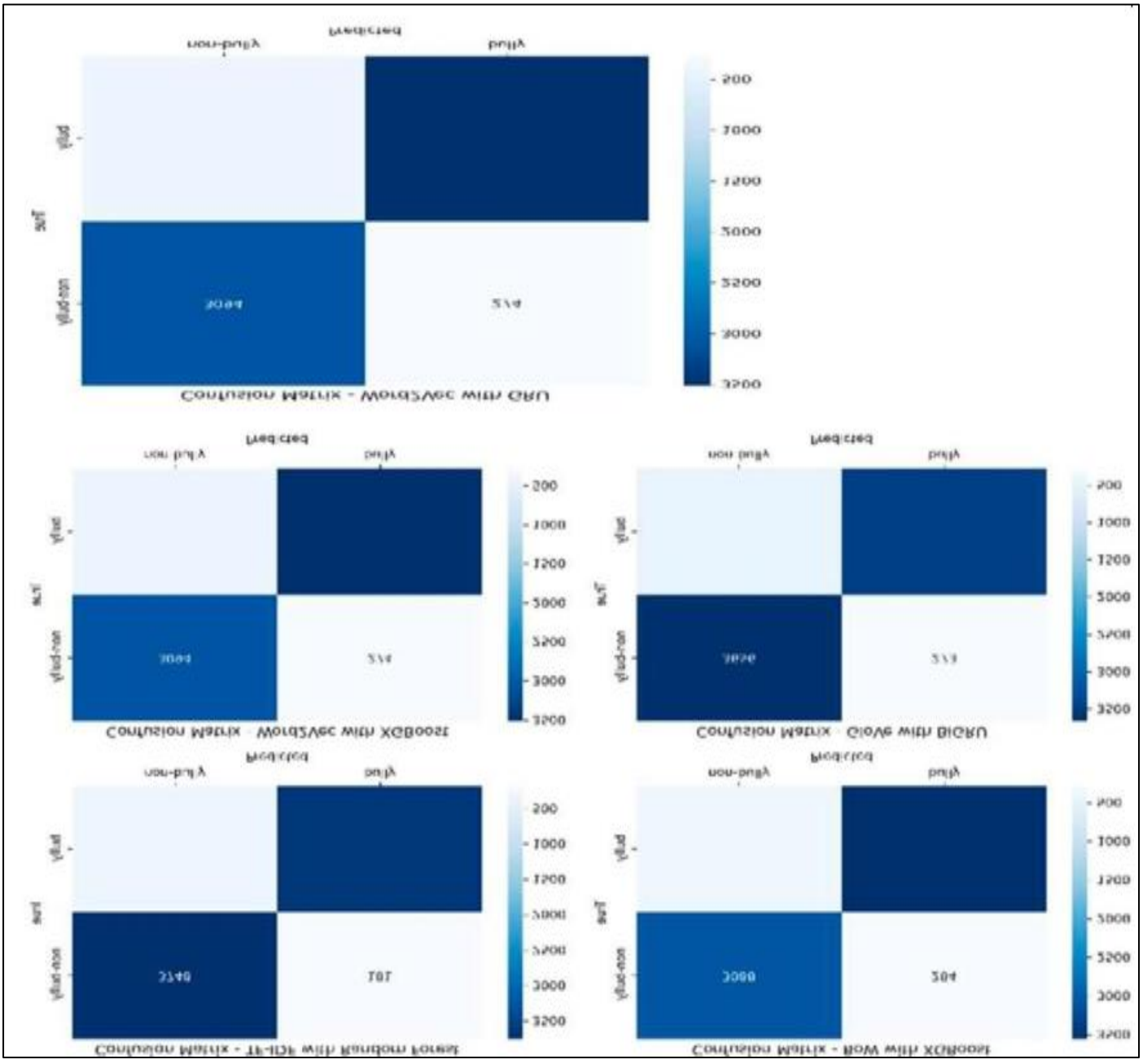

Fig 2 Confusion Matrices for Better- Performing Mod

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

These performance evaluations confirm the strength of combining suitable text embeddings with advanced classifiers to detect cyberbullying effectively. The combination of TF-IDF and Random Forest emerged as the most reliable overall.

- *Swear Words Dataset and Rating*

The Swear Words Dataset developed for this research consists of a carefully curated list of offensive terms commonly found on social media platforms. This list includes 428 swear words that were manually selected and rated to reflect the perceived level of offensiveness.

To ensure a balanced and inclusive rating process, individuals of various gender identities participated in assigning numerical severity scores to each word. These ratings were based on subjective perception, allowing the system to better capture the emotional weight and social impact of each term.

users	text	classification	rating	status
user 648	These so called Puerto Rican women always upli...	bully	4.0	Blocked
user 209	what's sad is that somewhere in Detroit @RealB...	bully	4.3	Blocked
user 885	i'm reading famous wattpad stories for inspira...	bully	4.7	Blocked
user 706	#Coon	bully	5.0	Blocked
user 217	You know, Rifftrax sure does a lot of jokes wh...	bully	1.0	Warning
user 545	Racist jokes, rape jokes, gay jokes, are all p...	bully	2.0	Warning

Fig 3 Sample Screenshot of the Rating System

This dataset plays a key role in enhancing the post-classification phase of the cyberbullying detection model. Once a comment is flagged as bullying, the presence of swear words is analyzed, and a severity score is calculated based on the ratings in the dataset. The average score is used to guide platform-specific responses—either issuing a warning or recommending account restriction.

Figure 3 provides a sample view from the rating system. Each entry displays the comment text, the associated severity score, and the system's response based on the set threshold.

By integrating this custom dataset into the classification pipeline, the model gains a more refined understanding of language nuances. This contributes to the overall reliability and fairness of the system in detecting and responding to abusive content.

V. CONCLUSION

This research addresses the growing concern of online bullying on digital platforms by introducing a robust detection system that combines both traditional machine learning and advanced deep learning techniques. Through comprehensive testing, Random Forest stood out as the most accurate model among ML methods, while Bi-LSTM and Bi-GRU demonstrated strong performance within the DL group.

A key strength of this study lies in its two-layered structure: the initial classification of content and the subsequent severity scoring using a curated swear words dataset. This additional layer not only identifies offensive language but also evaluates its intensity, enabling more informed and context-sensitive actions.

The curated list of 428 offensive terms, manually rated for severity, enhances the system's adaptability to varied user behavior and language usage patterns. By incorporating this dataset, the model gains a better grasp of subtle linguistic cues, improving its ability to distinguish between mild and severe bullying.

The real-world value of this work is demonstrated through the implementation of an automated moderation mechanism. When a message surpasses the predefined severity limit, the system proactively suggests blocking the user, whereas less harmful content triggers a warning. This makes the framework suitable for practical deployment on social platforms aiming to protect users, especially those frequently targeted like public figures and influencers.

In conclusion, the proposed system offers a balanced, scalable, and effective solution for detecting and managing cyberbullying, contributing to safer and more respectful digital communities.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

FUTURE WORKS

Looking ahead, future efforts will focus on expanding the scope and depth of the existing cyberbullying detection system. One major direction involves enriching the dataset by incorporating a wider variety of offensive expressions, including slang, abbreviations, and evolving online language. This will allow the model to adapt more effectively to the constantly changing dynamics of digital communication.

Another key area of improvement is the refinement of the rating mechanism. Introducing a more granular scoring scale or context-aware filtering can lead to even more precise severity assessment, ensuring that subtle but harmful messages are not overlooked.

Additionally, integrating advanced natural language processing techniques—such as transformer- based models like BERT or RoBERTa—may further enhance the system's ability to understand tone, sarcasm, and deeper contextual meanings in user comments.

Long-term, continuous adaptation will be essential. By monitoring shifts in language usage and user behavior patterns, the detection framework can be updated to stay current with emerging trends and threats. This responsiveness will help maintain the effectiveness of cyberbullying moderation systems as online discourse continues to evolve.

REFERENCES

[1]. J. Doe, R. Johnson, and S. Williams, "Understanding online bullying: A holistic Review," Journal of Cybersecurity, vol. 10, no. 3, pp. 123-145, 2023.

[2]. Smith, B. Wilson, and C. Brown, "Online Harassment of Public Figures: A Case Study on Celebrity online bullying," in International Conference on Cybersecurity, pp. 56-67, 2023.

[3]. K. Brown, J. Smith, L. Johnson, and M. Garcia, "Social Media Trends Post-Pandemic," Journal of Digital Communication and Social Media, vol. 8, no. 2, pp. 201-215, 2023.

[4]. D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, "Applied logistic regression," John Wiley & Sons, 2013.

[5]. L. Breiman, "Random forests," ML, vol. 45, no. 1, pp. 5-32, 2001.

[6]. V. Vapnik, "The nature of statistical learning theory," Springer Science & Business Media, 1995.

[7]. T. Chen and C. Guestrin, "Xgboost: A scalable and portable parallelized tree boosting framework," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[8]. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[9]. Graves, M. Schuster, and J. Schmidhuber, "Bidirectional recurrent neural networks," IEEE transactions on speech and audio processing, vol. 13, no. 5, pp. 793-801, 2005.

[10]. K. Cho, B. Merrienboer, S. Gulcehre, D. Bahdanau, F. Boureau, Y. Bengio, and C. Baroni, "Learning phrase representations using recurrent neural networks," arXiv preprint arXiv:1406.1054, 2014.

[11]. J. Chung, S. Gulcehre, K. Cho, and Y. Bengio, "Gated recurrent units with applications to language modeling and sentiment analysis," arXiv preprint arXiv:1412.3900, 2015.

[12]. R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring Hate Speech Detection in Multimodal Publications," Computational Linguistics Review, vol. 20, no. 3, pp. 112-129, 2019.

[13]. K. Reynolds, A. Kontostathis, and L. Edwards, "Using ML to detect online bullying," Journal of Social Networking, vol. 5, no. 2, pp. 112-125, 2011.

[14]. S. Mitchell et al., "Cyber-aggression, online bullying, and Cyber-grooming: A Survey and Research Challenges," J. Cybersecurity Res., vol. 5, no. 2, pp. 123-145, 2019.

[15]. J. Salminen et al., "Developing an online hate classifier for multiple online social networks," Journal of Computational Social Science, vol. 3, no. 2, pp. 333-348, 2020.

[16]. P. Mishra, H. Yannakoudakis, and E. Shutova, "Tackling Online Abuse: A Survey of Automated Abuse Detection Methods," J. Nat. Lang. Process. Comput. Linguist., vol. 17, no. 3, pp. 215-238, 2019.

[17]. N. Mehendale, K. Shah, C. Phadtare, and K. Rajpara, "Cyber Bullying Detection for Hindi- English Language Using

[18]. ML," SSRN, 2022. [Online]. Available: https://ssrn.com/abstract=4116143.

[19]. S. Neelakandan, M. Sridevi, S. Chandrasekaran, K. Murugeswari, A. K. Singh Pundir, R. Sridevi, and T. Bheema Lingaiah, "DL Approaches for online bullying Detection and Classification on Social Media," Hindawi Computational Intelligence and Neuroscience, vol. 2022, article ID 2163458, pp. 1-13, Jun. 11, 2022. DOI: 10.1155/2022/2163458.

[20]. B. A. Talpur and D. O'Sullivan, "online bullying severity detection: a ML approach," PLoS One, vol. 15,no. 10, Article ID e0240924, 2020.

[21]. N. Yuvaraj, V. Chang, B. Gobinathan et al., "Automatic detection of online bullying using multi-feature based artificial intelligence with deep decision tree classification," Computers & Electrical Engineering, vol. 92, Article ID 107186, 2021.

[22]. Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "ML and feature engineeringbased study into sarcasm and irony classification with application to online bullying detection," Information Processing & Management, vol. 58, no. 4, Article ID102600, 2021.

[23]. C. Van Hee et al., "Automatic detection of online bullying in social media text," PLoS ONE, vol. 13, no. 10, e0203794, 2018.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jul233

[24]. Perera and P. Fernando, "Accurate online bullying Detection and Prevention on Social Media," Procedia Computer Science, vol. 181, pp. 605-611, 2021.
[25]. D. Yin, Z. Xue, and L. Hong, “Detection of Harassment on Web 2.0,” p. 8, 2019.
[26]. S. O. Sood, E. F. Churchill, and J. Antin, “Automatic identification of personal insults on social news sites,” J. Am. Soc. Inf. Sci. Technol., vol. 63, no. 2, pp. 270–285, Feb. 2012, doi: 10.1002/ asi.21690.
[27]. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, “Identification and characterization of online bullying dynamics in an online social network,” in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM ’15, Paris, France, 2015, pp. 280–285, doi: 10.1145/2808797.2809398.
[28]. V. S. Chavan and Shylaja S S, “ML approach for detection of cyber-aggressive comments by peers on social media network,” in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, Aug. 2015, pp. 2354–2358, doi: 10.1109/ICACCI.2015.7275970.
[29]. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proceedings of the NAACL Student Research Workshop, pp. 88-93, 2016.
[30]. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proceedings of the 11th International AAAI Conference on Web and Social Media,
[31]. pp. 512-515, 2017. [Online]. Available: https://github.com/t- davidson/hate-speech-and-offensive- language.