International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Preventing Digital Harm through AI: A Hybrid Model for Detecting Cyber-Bullying and Emotional Risk on Social Media

Balram Singh Yadav[1], Dr. Saurabh Sharma[2]

Submitted: 02/11/2024 **Revised:** 18/12/2024 **Accepted:** 28/12/2024

Abstract: The rise of social media has introduced a new form of psychological toxicity cyber-bullying, which significantly threatens mental well-being, particularly among adolescents and young adults. The emotional consequences of online harassment, such as anxiety, depression, and suicidal ideation, have increasingly been recognized as public health risks. The Hybrid Supervised Cyber-bully Detection System (HS-CBDS) integrates multiple machine learning techniques to enhance classification accuracy and reliability. The system incorporates comprehensive text preprocessing including normalization, tokenization, stopword removal, and lemmatization followed by feature engineering using TF-IDF, Bag of Words (BoW), sentiment polarity, and text length. A rule-based filter checks for offensive terms using a curated lexicon. The model employs supervised learning classifiers Linear SVC, Logistic Regression, and Random Forest optimized via hyperparameter tuning and GridSearchCV. In the HS-CBDS system, ensemble learning is implemented by combining predictions from five different classifiers LinearSVC, Logistic Regression, Random Forest, Decision Tree, and MLPClassifier. Each model's prediction is assigned an equal weight (0.2), and the final decision is made by averaging these outputs and rounding the result. Experimental results demonstrate that HS-CBDS outperforms individual classifiers, achieving an accuracy of 99.33%, significantly outperforming individual models, including ANN (98.17%) and Random Forest (97.99%). Evaluation metrics such as precision (99.34%), recall (99.33%), F1-score (99.33%), and ROC-AUC (0.9932) further validate the robustness and reliability of the system. HS-CBDS demonstrates a scalable, interpretable, and high-performing solution suitable for real-time cyber-bullying detection across social media and educational platforms, contributing effectively to a safer digital environment.

Keyword*: cyber-bully, social Networking sites, social media, cybercrime, Harassment.*

1. INTRODUCTION

Now a day's technology plays an important role in our daily life. In last few years due to the availability of low cost of internet not only revolutionized our way of life but also the way we exchange ideas and information online. More people are able to use the internet and are spending a lot of their day time on social networking sites. But this increase in internet involvement has unintentionally led to the disturbing problem of cyber-bullying.

Cyber-bullying is when someone intentionally and aggressively targets someone who can't protect themselves from these attacks using digital communication. Cyber-bully can be done by an individual or a group of people. [1] The effect of cyber-bullying on victims separates it from traditional forms of bullying. Cyber-bullying causes deeper wounds, hurting victims psychologically and emotionally, whereas traditional bullying often results in physical and emotional harm.

It is crucial to set up monitoring and detection systems for potentially dangerous online conduct given the adverse impacts of cyber-bullying. This is the reason that creating a system to detect bullying texts or messages on social media becomes crucial in order to reduce such terrible incidents. This topic focuses on developing and implementing a workable solution that uses machine learning algorithms to identify abusive or harassing texts in the world of digital communications. Aggressive behavior when using information and communication channels, such as the internet, gaming consoles, or smartphones, is known as cyber victimization. People are thus

1Research scholar, Sant Baba Bhag Singh University, Jalandhar PUNJAB, INDIA
Email id- balram@davcollegeasr.org
2Associate Professor, Sant Baba Bhag Singh University, Jalandhar PUNJAB, INDIA
Email id- cybersense99@gmail.com

made victims of crimes, harassment, trolling, stalking, bullying, and violence in the online community [2]. Cybercrimes, including cyber-bullying, are on the rise in India, a country that is rapidly growing in cyberspace. With more than 33% of youngsters reporting having experienced online abuse, India has the greatest rate of this type of harassment. According to the latest data from the National Crime Records Bureau (NCRB), cybercrime cases in India have continued to rise. In 2022, a total of 65,893 cybercrime cases were registered, marking a 24.4% increase from the 52,974 cases reported in 2021. The crime rate under this category increased from 3.9 per lakh population in 2021 to 4.8 in 2022. Notably, 64.8% of these cases were related to fraud, followed by extortion (5.5%) and sexual exploitation (5.2%). This issue draws attention to the risks that strangers and imposters pose, especially to young people. Cyber victimization is characterized by aggressive acts such as sending abusive emails and messages. Harassing emails and messages, sexist comments, sharing or posting embarrassing images or videos, online threats, intimidation, and blackmail are examples of aggressive activities that constitute cyber victimization.[3]

2. RELATED WORK

There are many approaches that propose systems by researcher which can detect cyber-bullying messages automatically with high accuracy. Author Nandhini et al. [4] built a model utilizing data from MySpace.com that achieved 91% accuracy using the Naive Bayes machine learning approach. They then proposed a model that employs genetic operations (FuzGen) and the Naive Bayes classifier to attain 87% accuracy. Another approach by Romsaiyud et al. [5] they enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering and by this approach they achieved 95.79% accuracy on datasets from Slashdot, Kongregate, and MySpace.

Romsaiyud et al. [6] investigate word extraction using the Naive Bayes classifier and clustering of loaded patterns. The algorithm divided the datasets from Slasddot, Kyspace and Kongregate. And achieve 95.79% accuracy. Furthermore, Bunchanan et al. [7] .'s approach made use of a dataset that was manually categorized from War of Tanks game discussion. The results of study were inferior to those of the human classified dataset when compared to simple Naive classification that incorporates sentiment ana lysis as a feature. Additionally, Isa et al.h [8] suggested a method in which they obtained their dataset from Kaggle and employed two classifiers, Naive Bayes and SVM. The Naive Bayes classifier had an average accuracy of 92.81 percent compared to the 97.11 percent of SVM with poly kernel's. However, The dataset used by Dinakar et al.[9] to identify explicit bullying language relating to sexuality, IQ and ethnicity and culture came from the YouTube comment area.

When SVM and Naive Bayes classifiers were used, SVM had an accuracy of 66% and Naive Bayes had an accuracy of 63%. Di Capua et al. [10] proposed a brand-new method for detecting cyber-bullying using an unsupervised approach Growing Hierarchical SOMs. On FormSpring get 73% accuracy, YouTube 69% accuracy and 72% accuracy on twitter. A model to detect cyber-bullying developed by Haidar et al. [11] Naive Bayes, and SVM on Arabic language and achieved 90.85% precision and SVM achieved 94.1% as precision.

Zhao et al. [12] provided a framework for identifying cyber-bullying; they utilised word embedding to create a list of pre-defined undesirable phrases and gave weight’s to acquire bullying traits; they get result 79.4% using SVM as their primary classifier. Parime et al. [13], manually annotated on the dataset from MySpace, it, and used the SVM Classifier to classify it, proposed a different strategy. Chen et al. [14] suggested an unique feature extraction technique called Lexical Syntactic Feature and obtained 77.8% recall and 77.9% accuracy using SVM as their classifier. Ting et al. [15] established a method based on Social network mining method they included elements such as sentiments and Social network analysis measures, and their data came from social media. Seven experiments were conducted, and the precision and recall were nearly 97% and 71%, respectively. In addition, a framework named SICD was introduced by Harsh Dani et al. [16] that uses KNN for categorization. Their final grades were 0.7539 AUC and 0.6105 F1. Hee et al. [17] report the collection and fine-grained annotation of an English and Dutch cyber-bullying corpus, which was done in order to demonstrate the viability of automatic cyber bullying detection. Additionally, they carried out a number of binary classification trials. They examine the application of highly feature-rich

linear support vector machines to locate posts that are connected to cyber-bullying. 64% for English and 61% for Dutch are their respective scores. Theyazn H. H. Aldhyani et al.[18] using Convolutional neural networks combined with a single BiLSTM and bidirectional long short-term memory networks (CNN-BiLSTM) to detect cyber-bully texts with 99% accuracy. Pradeep Kumar & Fenish Umeshbhai [19] use the deep learning-based convolutional neural network model and got 89% accuracy for detecting the cyber-bully post. Dewani A et al.[20] use RNN-LSTM, RNN-BiLSTM models for detecting the cyber-bully method in Urdu-Roman language and got accuracy 85.5 and 85% whereas F1 score was 0.7 and 0.67 respectively. Bharti, S. et al.[21] use deep learning algorithms for word embedding technique on 35,787 with BLSTM and achieve outcomes on the dataset with 92.60% accuracy, 96.60% precision, and 94.20% F1 measure, respectively. Hani et al.,[22] in 2019 proposes a supervised machine learning approach for detecting and preventing cyber-bullying using classifiers to train and recognize on cyber-bullying dataset. Results show that Neural Network performs better and achieves accuracy of 92.8% and SVM achieves 90.3. Where NN outperforms other classifiers of similar work on the same dataset.

3. RESEARCH METHODOLOGY

3.1 Dataset

This dataset was taken from kaggle.com [23]. Data set contain total 99990 comment and out of which 49990 comment are cyber-bully comment label as "1" and remaining 50000 comments are non-cyber-bully messages label as "0". The texts or comments were divided in to two types as follows:

- **Non cyber-bullying Text**: This type of comments or posts is non-cyber-bullying or positive comments. For example, the comment like "You look super sexy and beautiful." is positive and non-cyber-bullying comments. These comment label as "0".

- **Cyber-bullying Text:** This type of comments or posts is cyber-bullying or negative comments. For example, the comment likes "you bitch, I will kill you" Is positive and cyber-bullying comments. These comment label as "1".

Text preprocessing is a significant stage of cyber-bullying identification where raw text data is cleaned and normalized to prepare it for successful analysis. Text normalization starts the process by making all the characters in lowercase so that uniformity is preserved. After this, noise is removed by stripping unwanted things like URLs, social media tags (e.g., @username), hashtags, and special characters while leaving alphabetic content only behind. This is followed by tokenization and stopword elimination, which split the text into words (tokens) and eliminate frequent words such as "and" and "the" that do not add significant information. To further process the text, lemmatization is used, which converts words to their root forms (e.g., "running" to "run"), enhancing consistency and minimizing redundancy. Lastly, the purified tokens are put together in a formatted text manner so that the data becomes machine learning model-ready. This organized preprocessing boosts the model's performance in identifying cyber-bullying messages more accurately and effectively. Fig 1 represent proportion of cyber-bullying and Non cyber-bullying message.

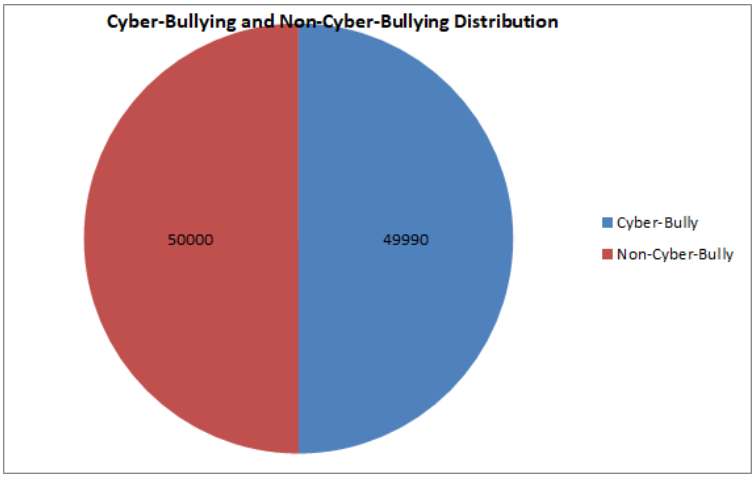

Fig 1 cyber-bullying vs. Non cyber-bullying Distribution

3.2 PROPOSED METHODOLOGY

The Hybrid Supervised Cyber-bully Detection System (HS-CBDS) is designed to analyze social media text, extract key features, and classify messages as cyber-bullying or non-cyber-bullying. This hybrid approach enhances detection accuracy by integrating text-based attributes with supplementary features.

The system comprises several core components:

1. **Text Preprocessing**: Social media texts are cleaned and normalized using advanced NLP techniques to prepare them for analysis.

2. **Feature Engineering**: Extracts meaningful features such as **TF-IDF**, **Bag of Words (BoW)**, **sentiment scores**, and **text length**, which are critical for classification.

3. **Model Training & Hyperparameter Tuning**: Five machine learning models — **LinearSVC**, **Logistic Regression**, **Random Forest**, **Decision Tree**, and **MLPClassifier (ANN)** — are trained and optimized to improve performance.

4. **Memory-Efficient Design**: Techniques like **sparse matrices** and **garbage collection** are employed to manage large datasets efficiently.

5. **Ensemble Learning**: Uses a **weighted voting strategy** to combine predictions from all classifiers, significantly improving accuracy and reliability.

This robust and scalable design enables HS-CBDS to effectively tackle cyber-bullying detection challenges in social media. The overall workflow is illustrated in **Fig. 2**.

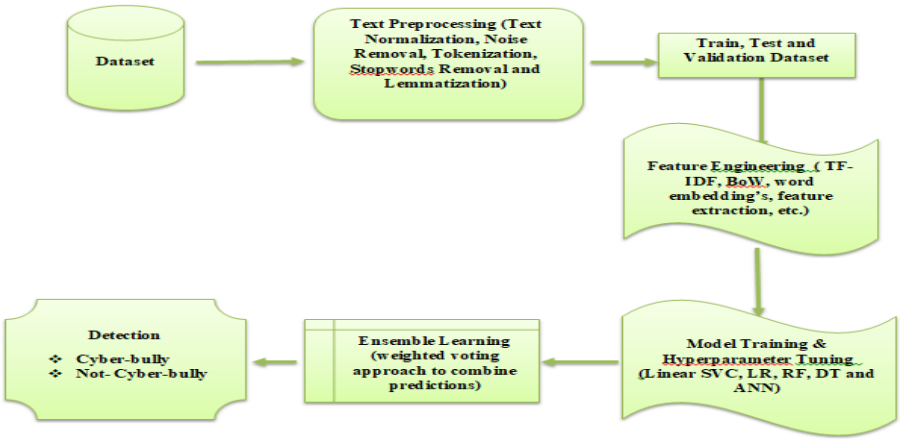

Fig. 2 Architecture of the proposed HS-CBDS

The architecture of the Hybrid Supervised Cyber-bully Detection System (HS-CBDS) is illustrated in architecture. It begins with essential Python imports for data processing, natural language processing (NLP), machine learning, and evaluation. Key libraries include pandas, numpy, re, nltk, and sklearn. Before processing, required NLTK resources like stopwords, punkt, and wordnet are downloaded to support text preprocessing.

A predefined offensive word list derived from survey data acts as an additional indicator of abusive content. Words like “idiot”, “bc”, “bsdk”, “stupid” “ugly” etc. help flag potentially harmful language during message analysis. The core preprocessing function, preprocess_text, converts all text to lowercase, removes non-alphabetic characters using regular expressions, and applies tokenization, stopword removal, and lemmatization. This standardizes and simplifies the text for better model performance. An additional function, add_features, computes sentiment polarity (via TextBlob) and text length. These supplementary features help models better understand emotional tone and message structure key indicators of cyberbullying. Another rule-based function, check_bag_of_words, scans input messages for any offensive terms, triggering a warning if any are found. After preprocessing, the dataset is loaded from a CSV file with labeled data (0 = non-cyberbullying, 1 = cyber-bullying). Text features are extracted using TF-IDF vectorization (unigrams and bigrams) with up to 5000 features, discarding overly rare or common words using min_df=2 and max_df=0.95. Numeric features like sentiment and text length are scaled with StandardScaler and then combined with TF-IDF vectors into a unified feature matrix using

scipy.sparse.hstack. The dataset is split into training and test sets (80/20 split), and five machine learning classifiers are trained: LinearSVC, Logistic Regression, Random Forest, Decision Tree, and MLPClassifier (ANN). Each model undergoes hyperparameter tuning using GridSearchCV with 3-fold cross-validation to find optimal configurations. To boost performance and reliability, ensemble learning is applied. Each model's predictions are weighted equally (0.2) and averaged to produce the final classification. This approach balances individual model strengths and reduces the risk of bias or overfitting. Evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC, providing a comprehensive view of model effectiveness. Finally, the system includes an interactive interface where users can input a message. The message is preprocessed, feature-extracted, and passed through the same trained models. The ensemble then predicts whether the message is cyber-bullying. If offensive terms are found or the ensemble result is positive, a warning is displayed; otherwise, the message is deemed safe.

The modular, interpretable, and scalable design of HS-CBDS makes it suitable for deployment in real-time moderation systems across social media and educational platforms. Its hybrid nature ensures high accuracy while maintaining transparency for end-users.

4. PERFORMANCE AND EVALUATION

The performance evaluation of various machine learning classifiers for cyber-bullying detection reveals that the Hybrid Supervised Cyber-bully Detection System (HS-CBDS) significantly outperforms individual models in terms of accuracy, precision, recall, and F1-score. Table 2 shows the Comparison of HS-CBDS with traditional algorithm. Amongst the traditional models, ANN was the best with a maximum accuracy of 98.17%, followed by RF with 97.99% and Decision Tree, Logistic Regression (LR), SVM, NB with 88.63%, 86.48%, 87.27% , 88.51% accuracy, respectively. The k-Nearest Neighbors (KNN) model, however, did less an accuracy of 82.19%. Despite these performances, the proposed HS-CBDS (Hybrid Supervised Cyber-bullying Detection System) significantly outperformed with all existing models, achieving an impressive accuracy of 99.33% along with high precision (99.34%), recall (99.33%), and F1-score (99.33%). This indicates that the novel hybrid approach provides a more effective solution for detecting cyber-bullying messages on social media. The accuracy, precision, recall and F1 score results are shown in Fig. 3, Fig. 4, Fig. 5 and Fig. 6 respectively. Additionally, the high ROC-AUC score of 0.9932 in Fig. 7. By combining the predictive power of diverse supervised learning algorithms, the proposed system minimizes misclassification errors and enhances reliability in detecting cyber-bullying messages. This significant advancement in cyber-bullying detection offers a more robust and scalable solution for identifying harmful content on social media platforms, ultimately contributing to a safer online environment.

Table 1 Comparison of **HS-CBDS** with traditional algorithm

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1 Score**
SVM	0.8727	0.8895	0.9662	0.9263
DT	0.8851	0.8852	0.8852	0.8851
KNN	0.8219	0.8261	0.8116	0.8188
RF	0.9799	0.9852	0.9742	0.9797
LR	0.8648	0.8791	0.9700	0.9223
NB	0.8851	0.8377	0.9436	0.8875
ANN	0.9817	0.9821	0.9837	0.9817
HS-CBDS(Proposed)	**0.9933**	**0.9934**	**0.9933**	**0.9933**

The superior performance of the proposed model suggests that combining multiple supervised learning techniques enhances classification accuracy and reliability, making it a promising advancement in cyber-bullying detection.

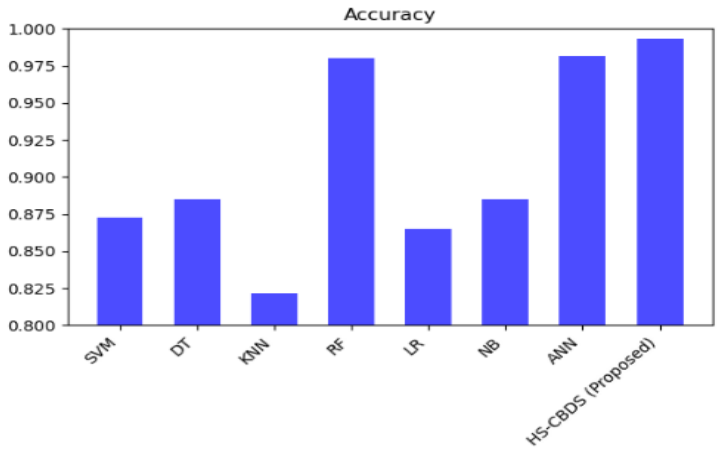

Fig. 3

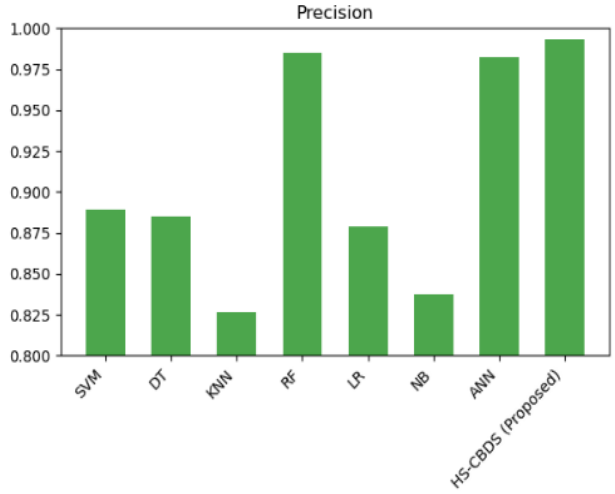

Fig. 4

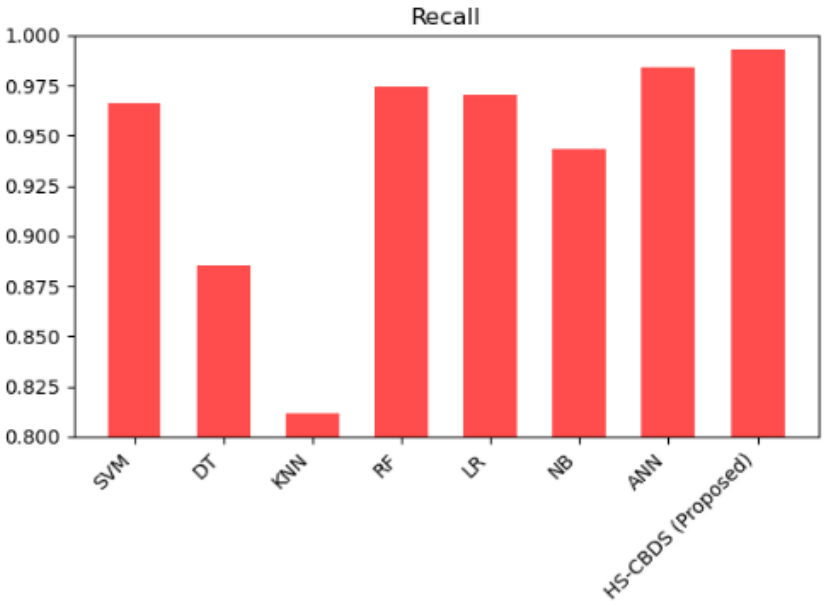

Fig. 5

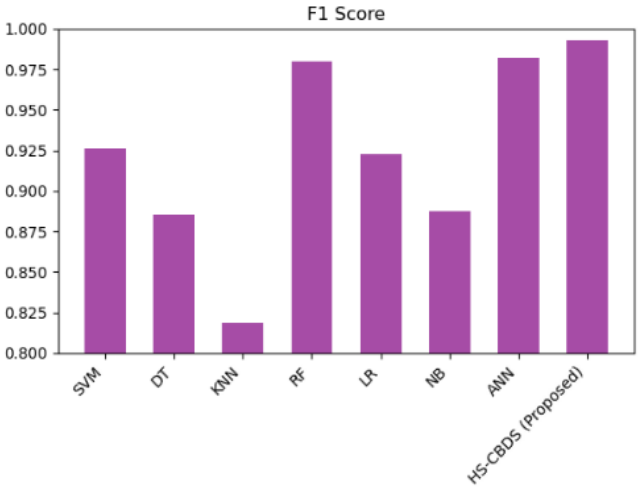

Fig. 6

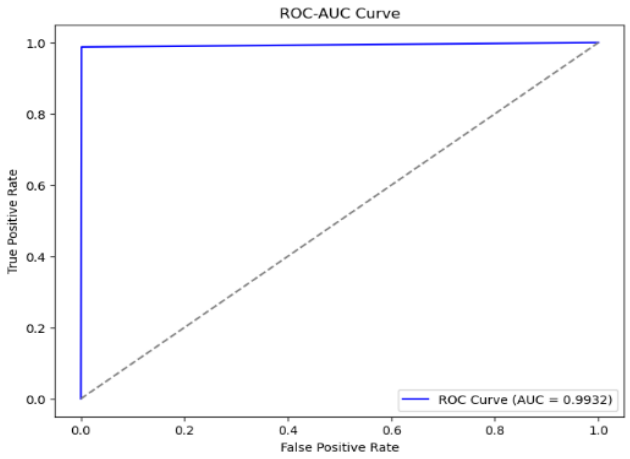

Fig. 7

The superior performance of HS-CBDS highlights its ability to enhance prediction reliability, reduce misclassification rates, and provide a more robust solution for cyber-bullying detection on social media. These results demonstrate that hybrid supervised learning techniques can effectively improve classification performance compared to traditional individual models.

5. CONCLUSION

This paper introduces the Hybrid Supervised Cyber-bully Detection System (HS-CBDS), a new machine learning-based method for social media cyber-bullying detection. By combining Linear SVC, Logistic Regression, and Random Forest in a weighted voting ensemble, the model greatly enhances classification accuracy, precision, recall, and F1-score. The HS-CBDS model attains an accuracy of **99.33**%, which surpasses conventional machine learning models including SVM, Decision Tree, KNN, RF, LR, and ANN. Success for the hybrid system lies in wide text preprocessing, feature engineering, hyperparameter, and memory-optimized design to make the model capable of discerning cyber-bullying messages from non-cyber-bullying ones. Weighted voting mechanism helps in perfectly balancing false negatives and false positives and providing enhanced reliability in detection against cyber-bullying.

6. FUTURE ENHANCEMENT

Future research may expand this framework to include by integrating real-time monitoring capabilities to detect cyber-bullying instantly as messages are posted. Multilingual support can be added to cover regional and global languages, increasing its applicability across diverse user bases. Incorporating deep learning models like BERT or LSTM could further improve contextual understanding and classification accuracy. The system can also evolve to include user behavior profiling and emotional impact analysis for better

risk assessment. Context-aware analysis, dynamic offensive word lists, and explainable AI techniques will enhance adaptability and transparency. Additionally, mobile app integration and partnerships with social media platforms can help deploy the system at scale, contributing to a safer online digital environment.

7. DECLARATIONS

Funding and/or Conflicts of interests/Competing interests

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. The authors also have no relevant financial or non-financial interests to disclose.

8. REFERENCES

[1] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: Its nature and impact in secondary school pupils. Journal of child psychology and psychiatry, 49(4):376–385, 2008.

[2] Cyber victimization during the COVID-19 pandemic: a syndemic looming large. Shoib S, Philip S, Bista S, et al. Health Sci Rep. 2022;5:0. doi: 10.1002/hsr2.528.

[3] Cyberbullying among youth in developing countries: a qualitative systematic review with bibliometric analysis. Saif AN, Purbasha AE. Child Youth Serv Rev. 2023;146:1–10.

[4] B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015.

[5] B Sri Nandhini and JI Sheeba. Online social network bullying detection using intelligence techniques. Procedia Computer Science, 45:485–492, 2015.

[6] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. Automated cyberbullying detection using clustering appearance patterns. In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242– 247. IEEE, 2017.

[7] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security, 76:197–213, 2018.

[8] Sani Muhamad Isa, Livia Ashianti, et al. Cyberbullying classification using text mining. In Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017.

[9] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

[10] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised cyber bullying detection in social networks. In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016.

[11] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal, 2(6):275–284, 2017.

[12] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking, page 43. ACM, 2016.

[13] Sourabh Parime and Vaibhav Suri. Cyberbullying detection and prevention: Data mining and psychological perspective. In Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on, pages 1541–1547. IEEE, 2014.

[14] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pages 71–80. IEEE, 2012.

[15] I-Hsien Ting, Wun Sheng Liou, Dario Liberona, Shyue-Liang Wang, and Giovanny Mauricio Tarazona Bermudez. Towards the detection of cyberbullying based on social network mining techniques. In Behavioral, Economic,

Socio-cultural Computing (BESC), 2017 International Conference on, pages 1–2. IEEE, 2017.

[16] Harsh Dani, Jundong Li, and Huan Liu. Sentiment informed cyberbullying detection in social media. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 52– 67. Springer, 2017.

[17] Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, et al. (2018) Automatic detection of cyberbullying in social media text. PLOS ONE 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

[18] Aldhyani, T.H.H.; Al-Adhaileh, M.H.; Alsubari, S.N. Cyberbullying Identification System Based Deep Learning Algorithms. Electronics 2022, 11, 3273. https://doi.org/10.3390/electronics11203273

[19] Roy, P.K., Mali, F.U. Cyberbullying detection using deep transfer learning. Complex Intell. Syst. 8, 5449–5467 (2022). https://doi.org/10.1007/s40747-022-00772-z

[20] Dewani A, Memon MA, Bhatti S. Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. J Big Data. 2021;8(1):160. doi: 10.1186/s40537-021-00550-7. Epub 2021 Dec 22. PMID: 34956818; PMCID: PMC8693595.

[21] Bharti, S., Yadav, A.K., Kumar, M. and Yadav, D. (2022), "Cyberbullying detection from tweets using deep learning", Kybernetes, Vol. 51 No. 9, pp. 2695-2711. https://doi.org/10.1108/K-01-2021-0061

[22] Hani, J., Mohamed, N., Mostafa, A. E., Emad, Z., Amer, E., & Mohammed, A. (2019). Social Media Cyberbullying Detection using Machine Learning. International Journal of Advanced Computer Science and Applications, 10(5). https://doi.org/10.14569/ijacsa.2019.0100587

[23] https://www.kaggle.com/datasets/momo12341234/cyberbully-detection-dataset.