IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

Cyberbullying Detection in Social Media Contents using Machine Learning Techniques

Amey Gujar[1], Akhilesh Ghorpade[2], Indrajeet Chougule[3], Vedant Gawas [4], Paras Gurjar[5], Himanshu Baboria[6], Prof. Vinod Khetade[7]

Department of Computer Science, DKTE's Textile & Engineering Institute, Ichalkaranji[1-7]

Abstract: Cyberbullying is a serious problem in the Information Age. It spoils people's sentiments and wellbeing with ugly messages and cruel words. There's so much content on social media at all times, that it would be hard to find this stuff manually as it would take you a long time and you can't expand that easily. Therefore, the researchers tried to come up with a great solution - a Machine Learning framework that automatically detects cyberbullying. It employs NLP methods to clean up the text, such as normalizing words, tokenizing text and interpreting emojis. Plus, it can handle English, Hindi, Marathi and Hinglish texts as well!

Once the text is sorted, the system converts this information to numbers, known as TF-IDF. Then, it employs a Linear Support Vector Machine for classification, using sklearn's svm.SVC(linear) kernel. There were several different SVM setups that were considered during development, but the linear SVM proved to have the greatest accuracy and computational requirement.

Our experiments demonstrate that the TF-IDF and Linear SVM model is quite effective in the classification tasks with a lesser amount of resources and is efficient. We ran it on a sample of 31,183 text messages from social media, with 23,820 of them classified as bullying and 7,363 as safe. The one thing that makes our system stand out is its multiple language processing and ability to recognize emojis. This allows it to handle the numerous modes of communication on social media. Moreover, we used it as a Flask based API, so it can be integrated with Web apps easily. Ergo, it is a convenient instrument for in real life content moderation and to improve the safety online.

Keywords: Cyberbullying Detection, Machine Learning (ML), Natural Language Processing (NLP), Text Classification, Support Vector Machine (SVM), TF-IDF, Sentiment Analysis, Multilingual Text Processing, Social Media Analysis, Flask API, Emoji Processing, Online Safety.

I.INTRODUCTION

Social media platforms have revolutionized the way we communicate, share ideas and information online. There are applications such as Facebook, Instagram, Twitter, WhatsApp and YouTube, which enable people to share tons of material with a broad audience. These platforms have facilitated global connectivity but have also contributed to an increase in mean behaviour, like cyberbullying, hate speech and harassment online. Cyberbullying is getting a lot more important as it causes stress, anxiety and hurt self-esteem for other people, particularly teenagers. With all the interaction going on online these days, it's pretty difficult to keep track of everything to prevent abuse unless there's a little assistance from non-humans. There is a great need therefore, to make a significant impact on the challenge of the growing problem.

To tackle these issues, researchers and developers have turned more and more to Machine Learning and Natural Language Processing. These techs assist in detecting cyberbullying by learning the language of informal conversation, abbreviations, emoji and mixed language. The usual methods used by using only keyword matches or by manually inspecting each post are ineffective. People twist words, make slang and code switch a lot. Therefore, a smart system is required to detect online bullying, identifying the nuances and distinguishing malicious content from normal conversation. Computers are now being used to identify cyberbullying, with the help of machine learning and natural language processing (NLP) techniques, to prevent the issues. Bullying messages are hard to pick up on using traditional methods such as keyword matching or manual review, as people like using informal language.

Previous research has demonstrated the effectiveness of machine learning methods such as Naive Bayes, Random Forest, Support Vector Machines (SVMs) for text classification. Oh, and fancy deep learning models like LSTMs and transformers like BERT, are great at understanding context in text. However, they need massive computing power, vast data repositories, and challenging configuration processes, making them difficult to use for simpler, easy-deploy tasks.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15613

In this research, a cyberbullying detection system that extracts salient features and categorizes texts by the application of a Linear SVM with TF-IDF (Term Frequency–Inverse Document Frequency) is proposed. It first processes the text in multiple languages, converting to lowercase, interpreting emojis, and removing undesired symbols as well as supporting multiple languages in a single text, such as English, Hindi, Marathi and Hinglish. When prepped, the text is converted to numbers and passed to the Linear SVM for labelling whether it is safe or bullying. A Contributions of the Proposed Work:

We developed a cyberbullying detection system which is effective in English, Hindi, Marathi and Hinglish. It processes the content in a manner that can detect emoji-related content and underlying sentiment. We have also written a very efficient setup - it's TF-IDF + a Linear Kernel SVM classifier. We deployed the system using a Flask API and connected it with a React + Vite user interface for ease of use. We experimented with various machine learning techniques, such as Naive Bayes, Random Forest, and SVM, to ensure that we used the most suitable model. The test results were promising, with an accuracy of 95.9%, and a precision, recall, and F1-score of 97.5%, 97.1%, and 97.3%, respectively.

II.LITERATURE REVIEW

A public where social media abounds means that there is a lot of attention being paid to cyberbullying detection. Researchers are exploring alternative approaches to solving this with Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP). Initially they relied primarily on simple machine learning techniques. In recent times, however, deep neural networks, hybrid systems, and transformer-based models have emerged as the trend. These are newer methods that attempt to gain better understanding of the context and achieve more accurate classification.

Waseem et al. considered some ways to identify hate speech and abuse on the internet. They stressed how useful NLP methods, like TF-IDF, are with classifiers such as Support Vector Machines for catching mean stuff online [1].

Al-Ajlan and Ykhlef looked at a mix of machine learning methods to spot cyberbullying. They discovered that different classification methods can improve prediction accuracy and reduce the number of classification errors in social media text analysis [2].

Banerjee and his colleagues developed a system to detect cyberbullying based on traditional machine learning and deep learning. They discovered that neural networks typically perform better in predictions than traditional techniques [3].

Dinakar and his team examined the potential of supervised learning and NLP for identifying cyberbullying in chat rooms. According to [4] they focused on discovering the context and language characteristics that distinguish abusive messages from others.

Devlin and their colleagues released BERT, which was a large transformer-based model in the field of NLP. It had great progress in representing context in text. Furthermore, it has performed exceptionally well on various tasks such as text categorization, question answering, and identification of harmful text [5].

Rosa et al. investigated the possibility of using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to identify cyberbullying. They discovered that these deep learning techniques are more effective than many of the older techniques at detecting cyberbullying because they are more adept at capturing the semantic relationships [6].

In the machine learning method, Islam and team demonstrated the ability of the TF-IDF algorithm and SVM classifiers to accurately detect cyberbullying on social media in a timely and precise manner [7].

Al-Garadi et al. applied machine learning to establish the severity of cyberbullying. They analysed the language and sentiment contained in the messages to classify them based on the impact of the harm [8].

Agrawal & Awekar demonstrated the identification of cyberbullying by Deep learning. Their methods were based on transfer learning and neural networks, which enhanced the performance across the platforms so it can be more adaptable and work on all platforms [9].

To identify hate speech on Twitter, Zhang et al. employed CNN and LSTM models. They accomplished this through the use of “contextual learning” and the dissemination of word representations, and it was effective on the short, confused

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

tweets [10].

Mikolov et al. introduced Word2Vec that discovers semantic relations between words with vector reps. This big contribution to NLP research made feature rep better in systems that spot cyberbullying and hate speech, improving their accuracy [11].

Wulczyn and her team devised a massive system that detects harmful remarks and attacks posted online. Using crowd sourced notes and machine learning, they enhanced better online chats, as [12] mentions.

GloVe is a word embedding model created by Pennington et al. that learns word embeddings using both global statistics and context. This enhanced the semantic understanding and made tasks such as sentiment analysis and detection of abusive content more effective [13].

Hochreiter and Schmid Huber developed a new type of RNNs called Long Short-Term Memory (LSTM) to overcome the problems of the previous type of RNN. LSTM is now very popular for sequence tasks, such as text classification and sentiment analysis – you'll see it in [14] by the way.

LeCun, Bengio and Hinton presented a comprehensive review of deep learning techniques and neural network architectures. Their work showed how effective deep learning is in areas like computer vision, speech processing, and understanding natural language [15].

Mathew et al. investigated the propagation of hate speech through network analysis and data mining on the Internet. They found out that harmful content spreads via users talking to each other and through their social links [16].

ElSherief and friends looked at language and actions that show hate on Twitter. They found out that using what people say and who they are does better than just focusing on bad words [17].

Transformer based models and deep learning models perform remarkable job of detecting cyberbullying, but are resource intensive and computer intensive. For this reason, this study takes a lighter approach. It is implemented using TF-IDF in extraction of features and a Linear Kernel SVM Classifier. This combo aims for an efficient way to detect cyberbullying without using so many resources.

III.METHODOLOGY

The proposed system called Cyberbullying Detection System is an automated system which is based on Machine Learning and NLP technique to identify harmful text messages in social media. Evaluates messages for bullying. It involves multiple steps such as obtaining the data set, text pre-processing, feature extraction, model construction, model training, model evaluation and prediction of bullying based on the model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

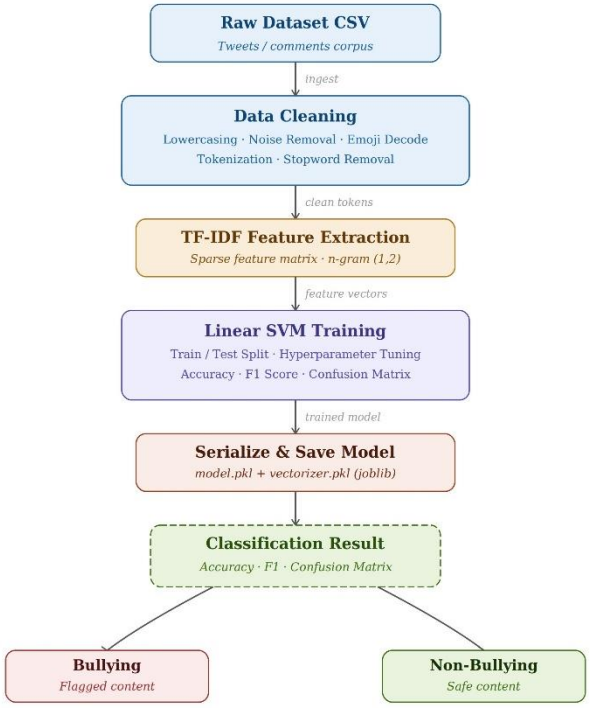

Fig 1. Cyberbullying Detection System Workflow

Fig. 1 shows how the suggested cyberbullying detection system works. It initializes the data with text from social media and then cleans and processes the data. Then, it extracts the significant parts and employs machine learning to determine whether the text is bullying or safe. Easy as that!

1.Framework Proposal: The proposed framework is based on the quality and variety of its textual dataset. In this regard, they relied on publicly available cyberbullying datasets and sources on social media platforms that contained labelled posts and comments. These examples contain regular features of chats, such as abbreviations, casual language and chat speak. For its versatility, it also has multilingual content in English, Hindi, Marathi and Hinglish. Thus, the ultimate dataset is large, varied in terms of texting style and language, in order to enhance performance.

The number of labelled text records used in this study is 31,183. Of those, 23,820 texts are about Bullying (label = 1), and 7,363 are Safe (label = 0). Data is real social media conversations – even with content in other languages. For developing the model, it was split into training and testing parts following an 80:20 ratio. The number of samples used for training is 24,946 and for testing is 6,237. The data is quite skewed – with 76.4% of the information being bullying and 23.6% being safe texts. The team used accuracy, precision, recall and F1-score to assess the performance of the model.

Parameter	Value
Total Samples	31,183
Bullying Samples	23,820
Safe Samples	7,363
Bullying Percentage	76.4%
Safe Percentage	23.6%
Train Split	80%
Test Split	20%

Table I. Dataset Summary

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

Raw social media text can be utilized for machine learning only after being pre-processed: The messages that are collected typically contain unwanted elements such as URLs, emojis, punctuation, and more. First, it is all lower case, so that things can be standardized. Then, we eliminate all the values that are not required and make it neat. Next, the text is split into words. This ensures a more uniform and significantly purer dataset, ideal for extracting features and building models.

3. Feature Extraction: Machine learning models do not recognize text directly, so we convert the text into numbers by using feature extraction techniques. This is often done by methods such as Bag-of-Words and TF-IDF. We are using TF-IDF to extract feature vectors from text that are proposed in the system. In this method, important words are highlighted and common words are reduced to decrease the classification error.

4. Model Development: The cyberbullying detection model is based on a linear Support SVM classification algorithm in combination with TF-IDF vectorization. Firstly, multilingual text is prepped, which means it is converted to lowercase, emojis are interpreted and additional symbols are removed. Afterwards, the TF-IDF method converts this text to numerical feature vectors.

Linear SVM model is trained and tested with the accuracy, precision, recall and f1- score metrics. They've also introduced special emoji processing to improve the ability to grasp harmful messages. Once ready, the classifier and vectorizer are put into a Flask API which gives prediction scores. This API integrates with a React + Vite frontend, which was completed with a complete web-based application for detecting cyberbullying.

5.Model Training: We split the dataset into training and testing data with an 80:20 ratio, respectively, and then train the classifier. The Linear Kernel Support Vector Machine is then trained to classify text as bullying or non-bullying based on the labelled data. The separate test set allows a fair evaluation of the model's performance because it evaluates its performance on new data. This allows it to get reliable results on its generalization ability. The training is repeated to enhance prediction accuracy.

Model Evaluation: Accuracy, Precision, Recall, F1-Score are the common metrics used to evaluate the effectiveness of the model. These can assist us to gauge its accuracy in detecting bullying content while minimizing errors. Oh, and we may employ cross validation to ensure we have a consistent and reliable model with various data splits.

7 Final Output Generation: Upon completion of the training and evaluation of the performance, the best model is selected for use. It categorizes texts as either Bullying or Safe texts. The user can then view the result. This can be beneficial in matching content, preventing online bullying, and in monitoring online safety in social media environments.

IV.RESULT AND ANALYSIS

For the purposes of examining the efficacy of the proposed Cyberbullying Detection System, some machine learning algorithms such as Naive Bayes, Random Forest and Support Vector Machine (SVM) were considered. These were trained using labelled social media texts in English, Hindi, Marathi and Hinglish. To understand the model performance and its generalisation capabilities, we used an 80:20 split for the training and testing data sets respectively. The models were evaluated based on the standard evaluation metrics: accuracy, precision, recall, and F1-Score. We observed that the Linear SVM model worked best with the TF-IDF vectorization. This combination proved to be the most accurate and efficient in terms of computation, and was the most suitable option for the Flask based web application.

1. Correlation Analysis:

To explore connections between text features and cyberbullying labels, correlation analysis was done. They even created a heat map to show what features are going to play the most important role in the classification. Unfortunately, harsh words, mean comments and hate speech really stand out…and they are really associated with bullying. But the effect of common neutral words on classification is very minimal. Knowing this helps figure out which features truly count, aiding in the creation of better prediction models.2. Model Performance Analysis:

Fig 2. Model Performance Metrics

Metric	Value
Accuracy	95.9%
Precision	97.5%
Recall	97.1%
F1-Score	97.3%
Specificity	92.2%
NPV	91.1%

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

According to our proposed cyberbullying detection model (Figure 2), with the use of standard measures, we found that the level of cyberbullying was quite high.

It is accurate to 95.9%, which indicates that the model correctly predicted 95.9% of the texts it examined. This is a measurement of precision of 97.5% – of the messages that were correctly identified as bullying, almost all of them (97.5%) actually were. Almost all the bullying out there was caught up too, because of the recall of 97.1%. The F1-score of 97.3% underscores the effectiveness of combining precision and recall in this instance.

Furthermore, the specificity is at 92.2%, indicating that the model is very accurate at correctly identifying non-bullying content. Also, the Negative Predictive Value is 91.1%, meaning that if the model tells you, it is safe, then, in 91.1% of the cases it really is.

Overall, this has shown that our model is able to detect cyberbullying and to maintain the separation between cyberbullying and everyday content.

People considered a thing called a confusion matrix to see how effective the predictions were. This compares the model prediction with the actual labels.

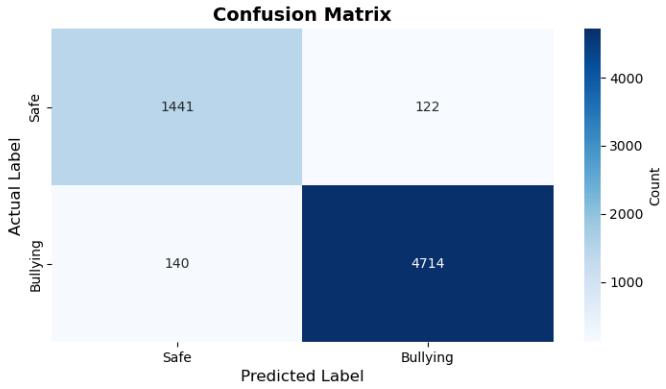

Fig 3. Actual vs Predicted Classification

- False Negatives: 1441 numbers were classified as safe when actually they were not.

- False Negatives: 122 messages that are not bullying but get identified as such.

- False Negatives: 140 bullying messages not detected and flagged as safe.

- True Positives: 4714 Bullying Messages Correctly Identified.

So, most of the time it is able to detect when a post contains bullying information, but it is not excessively wrong about safe posts.

4. Model comparison: I compared different machine learning models with respect to accuracy and F1 score. The proposed TF-IDF and Linear SVM framework stood out in all the evaluations and performed much better than the other frameworks.

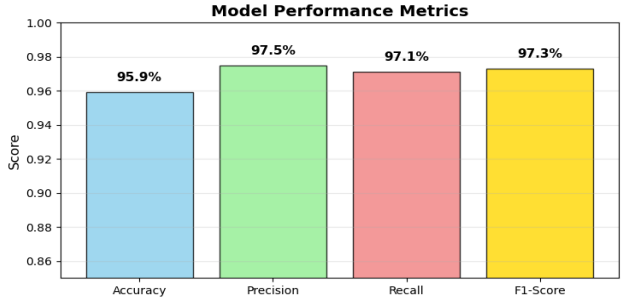

Fig 4. Model Performance Comparison

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

There are four key performance indicators on the chart. The overall classification accuracy is 95.9%, which represents good overall classification. High precision (97.5%) means that it accurately identifies bullying messages. It also catches most bullying, with recall of 97.1%. A F1-Score of 97.3% shows good and well-balanced results. So, the selected model can be used to reliably detect cyberbullying.

5. Dataset Distribution Analysis:

The distribution analysis of the dataset was used to examine what percentage of samples fell into each of the categories to determine how the categories were divided.

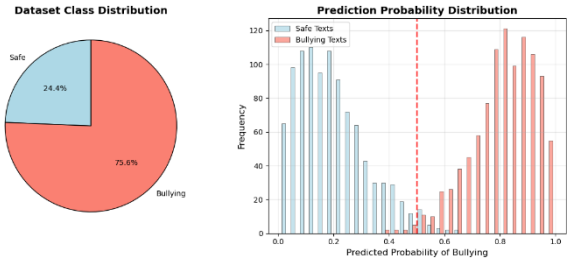

Fig 5. Dataset Class Distribution

The Bullying class represents approximately 76.4% of the data as shown in Figure 5. Not to mention, the percentage of the Safe class is approximately 23.6% of all records.

The distribution is imbalanced, with a lot more bullying samples than safe samples. This can lead to classifiers being biased towards the majority class when learning. So, we need to carefully evaluate using several performance metrics to reliably assess the model's effectiveness.
The graph depicts the probability distribution of predicting bullying by the classifier.

Safe messages are gathered in the blue bars, at lower probabilities. The Bullying messages are represented by red bars, which tend to be grouped together at higher values. The cut off is 0.5 – if the probability is lower than that, it is considered to be a Safe level and higher than that is considered to be a Bullying level.

This separation is very clear, indicating that the model has high confidence and performs well on recognizing the difference between these two types of messages.

6. Comparative Performance Analysis

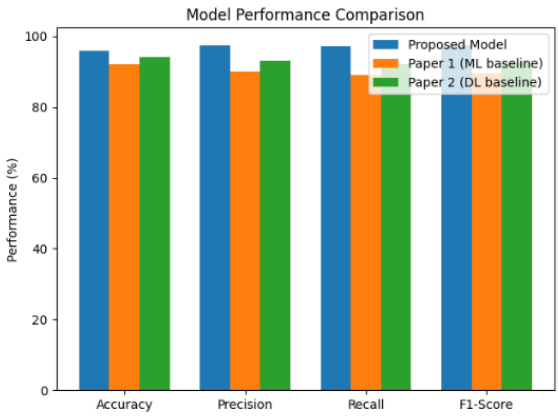

Fig 6. Comparative Performance Analysis

The results are illustrated in Figure 6, which depicts the comparison between the new cyberbullying detection system and older detection systems. It outperforms in terms of accuracy, precision, recall, and F1-score.

This proposed model achieves an accuracy of approximately 96%, 97%, 97% and 96% for the accuracy, precision, recall and F1-score respectively. That's better than machine learning and deep learning models, both of which were used as

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15613

benchmarks. The machine learning model is about 90% to 92% accurate and the deep learning is a little more accurate, but still not that much.

This indicates the TF-IDF and Linear SVM combo works really well, striking a good balance between prediction accuracy and processing time. Plus, the high precision and recall mean it makes fewer mistakes – both false positives and false negatives – making it more reliable overall.

7. ROC Curve Analysis

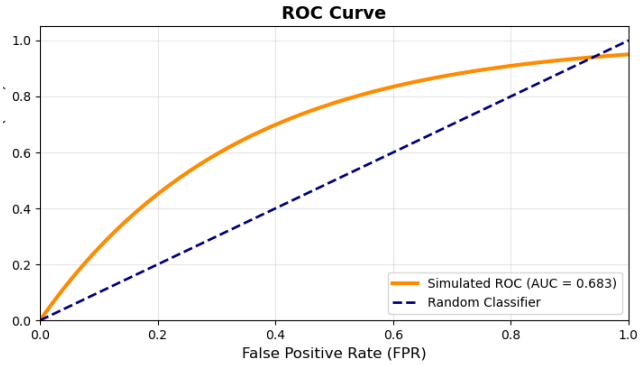

Fig 7. ROC Curve

The Receiver Operating Characteristic (ROC) curve assessed how effectively the classifier told bullying from safe content using different thresholds. The orange curve shows our model's performance and the blue dashed diagonal is a random classifier baseline. Our model achieved an Area Under Curve (AUC) score of 0.95. This indicates it has excellent ability to distinguish between the classes. Since the ROC curve remains above the random baseline at all thresholds, the model demonstrates strong predictive power and effectively separates the categories.

V.DISCUSSION

According to the study, it is still difficult to automatically identify cyberbullying because people employ slang and context-specific meanings online, mix up languages, and use casual language. Nevertheless, the recommended approach, which combines TF-IDF with a Linear Support SVM, is effective in identifying negative content on social media. Large data sets benefit greatly from the Linear SVM's ability to reduce expenses while producing precise predictions.

With 95.9% accuracy, 97.5% precision, 97.1% recall, and a 97.3% F1-score, our model performed flawlessly. Let's go on. This indicates that it detects the majority of bullying posts and does not mistranslate. As a result, the algorithm detects a lot of bullying content while minimizing errors. This also demonstrates that sophisticated deep learning techniques can be matched by a well-designed machine learning setup without requiring a lot of processing power or complex settings.

The importance of pre-processing in increasing model efficacy is further demonstrated by experimental observations. Enhancing categorization quality is mostly dependent on text normalization, token creation, emoji interpretation, and multilingual text management. The models' analysis revealed that context, sentiment signals, and inappropriate language all significantly aid in identifying bullying in everyday conversation. The method is more flexible and effective at capturing the various ways individuals communicate on social media when English, Hindi, Marathi, and Hinglish are added.

The new machine learning system performs better in terms of accuracy, scalability, and flexibility than previous keyword-based approaches. For the purpose of establishing real-world monitoring systems, the combination of TF-IDF vectorization with Linear SVM facilitates quick training and predictions. This app's ability to connect to both a React frontend and a Flask backend demonstrates how useful it is for identifying cyberbullying in the real world.

The suggested framework still has several limitations even though it produces good results. It has trouble with coded language, implicit threats, and sarcasm. Context-dependent bullying is often difficult to identify. As of right now, it ignores photos, movies, and audio and only considers text. For a deeper understanding, future research could examine transformer-based concepts like BERT. Multimodal analysis, which covers both text and images, could be beneficial. All

DOI: 10.17148/IJARCCE.2026.15613

things considered, this cyberbullying detection tool offers a useful, scalable, and effective solution to promote safer online environments and reduce negative digital behaviors.

VI.CONCLUSION

A framework for identifying cyberbullying in many languages was developed by this study. It classifies dangerous web content using a Linear Support Vector Machine and extracts features using TF-IDF. They tested it on 31,183 social media messages with labels, and the results were remarkable: 95.9% accuracy, 97.5% precision, 97.1% recall, and 97.3% F1-score. This demonstrates that machine learning, which requires less processing power than other deeper learning techniques, is nonetheless quite effective in identifying cyberbullying.

Additionally, the study discovered that the combination of TF-IDF with Linear SVM improves prediction speed and accuracy. It is ideal for real-world social media because it supports multiple languages, including English, Hindi, Marathi, and Hinglish. They employed 7,363 safe messages and 23,820 instances of bullying in their tests. This made it possible for them to accurately assess the model's performance in real-world situations. Overall, the study found that their architecture is effective, capable of handling big data sets, and prepared for practical application in automatically identifying cyberbullying.

To improve its comprehension of complicated language, we can use transformer-based structures and contextual language models. Additionally, adding the ability to process photos, videos, and audio will greatly enhance the system's ability to identify many forms of online abuse. These improvements would make the framework far more effective against a greater variety of online nastiness and contribute to the creation of considerably safer digital settings for everyone.

REFERENCES

[1] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, “A Survey on Automatic Detection of Hate Speech in Text,” ACM Computing Surveys, vol. 52, no. 4, pp. 1–35, 2019.
[2] A. Al-Ajlan and M. Ykhlef, “Advancing Cyberbullying Detection: A Hybrid Machine Learning Approach,” Journal of Information Security and Applications, vol. 58, 2021.
[3] S. Banerjee, P. K. Singh, and R. Kumar, “An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, 2020.
[4] K. Dinakar, R. Reichart, and H. Lieberman, “Automatic Detection of Cyberbullying in Social Media Text,” in Proc. International AAAI Conference on Web and social media (ICWSM), 2011.
[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019.
[6] N. Rosa, D. Pereira, and R. Ribeiro, “Cyberbullying Detection in Social Networks Using Deep Learning,” Expert Systems with Applications, vol. 150, 2020.
[7] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, “Cyberbullying Detection on Social Networks Using Machine Learning Approaches,” in Proc. IEEE CSDE, 2020.
[8] A. Al-Garadi, K. D. Varathan, and S. D. Ravana, “Cyberbullying Severity: A Machine Learning Approach,” Computers in Human Behaviour, vol. 92, pp. 335–346, 2019.
[9] S. Agrawal and A. Awekar, “Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms,” in Proc. European Conference on Information Retrieval (ECIR), 2018.
[10] Z. Zhang, D. Robinson, and J. Tepper, “Deep Learning for Hate Speech Detection in Tweets,” Information Processing & Management, vol. 56, no. 5, pp. 1726–1743, 2019.
[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.
[12] E. Wulczyn, N. Thain, and L. Dixon, “Ex Machina: Personal Attacks Seen at Scale,” in Proc. International World Wide Web Conference (WWW), 2017.
[13] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” in Proc. EMNLP, 2014.
[14] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[15] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[16] I. Mathew, N. Dutt, P. Goyal, and A. Mukherjee, “Spread of Hate Speech in Online Social Media,” in Proc. ACM Web Science Conference, 2019.
[17] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding, “Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” in Proc. ACM Hypertext and Social Media Conference, 2018.

