Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7546452

CYBERBULLYING DETECTION ON SOCIAL MEDIA: LEVERAGING TF-IDF AND LSTM FOR ROBUST CLASSIFICATION

Komati Lavanya
PG Scholor,
Computer Science Engineering,
Svs group of institutions,
Hanamakonda,TG,
chilukalavanya@gmail.com
k.Yakub Reddy
Assistant prof.,
Computer Science Engineering(DS),
Svs group of institutions,
Hanamakonda,TG,
yakubreddy1245@gmail.com

Abstract
This study investigates the detection of cyberbullying on social media using both traditional and advanced techniques. Specifically, it combines Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction with a Long Short-Term Memory (LSTM) deep learning model. The dataset, consisting of over 47,000 labeled tweets, was rigorously preprocessed to optimize feature extraction and minimize noise. The LSTM model, designed with a sequential architecture that incorporates embedding, LSTM, and dense layers, achieved an accuracy of 91%, demonstrating its effectiveness in capturing contextual information. In contrast, the TF-IDF method provided valuable interpretability, complementing the deep learning approach. The results demonstrate high performance; however, challenges persist in distinguishing between certain classes, particularly Class 3 and Class 4, due to their inherent complexity.
Keywords.*Cyber bullying, TF-IDF, LSTM, social media, deep learning.*

1. INTRODUCTION

The current era regards social media platforms as one of the most important components in society, as they still provide massive interaction, information exchange, and community establishment (Philipo, A. G., et al., 2024) [17]. According to the abstract, platforms such as Facebook, Twitter, Instagram, and TikTok have genuinely changed the way humans interact. The ubiquitous use of social media also brings a lot of problems, among which is the greatest risk of combating cyberbullying [2] (Fattahi, J., et al., 2024).

In particular, cyberbullying — the use of digital platforms to harass, threaten or demean others — erupted into a widespread problem, especially among youth. The repercussions are endless, affecting mental health, crippling confidence, and causing people to resort to isolation and, in some circumstances, self-harm or even death. Consequently, identifying and moderation of cyberbullying on social media has been an important research domain in the area of natural language processing (NLP) and machine learning (ML) (Saifullah et al., 2024)[10]; Kumar, K. M. K., et al., 2024)[15]. The anonymity and intrusion of social media make it an ideal breeding ground for cyberbullying. Bullying in the real-world often takes place in schools or workplaces, in short, limited environments, while cyberbullying has no geographical boundaries and can reach an extensive audience within minutes. Victims are significantly affected, as many experience anxiety, depression, and a fall in school or job performance. In addition, the pervasive nature of social media magnifies the hurt perpetrated through bullying messages, and it is almost impossible for the victim to avoid being targeted. Even though more and more people are becoming aware of the seriousness of this issue, there is still a long way to go to stop cyberbullying (Rezvani, N., et al., 2020) [19]. Social Media platforms have moderation systems by way of user reporting and automated filters that can identify abusive content, flags it, and remove it. But these are mostly passively and inadequate, as they depend on user reporting or prescriptive rule sets which do not capture the fluid and situational aspects of language. This signals a large potential to better understand and stop cyberbullying, with more adaptive, preventative methods. With recent developments in natural language processing, it is now possible to identify cyberbullying through the application of machine learning methods. These technologies allow for the analysis of large amounts of text data to find patterns and features suggestive of bullying behavior. So far, abusive language was detected using traditional NLP techniques like keyword matching, which specify a fixed sequence of characters, however, they do not take into account the nuances of human communication like sarcasm, context, cultural differences, and so forth.

In this regard, there are various machine learning algorithms that are developed to address such challenges, with deep learning models being the most critically reviewed algorithms among them (Aliyeva, Ç. O., and Yağanoğlu, M., 2024) [9]. These models, using large datasets and advanced architectures, are able to learn the complex patterns in the text and predict the presence of cyberbullying extremely well. Different ML techniques are available however, recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks (Sifath, S., et al. 2024)[7] have proven really well in a number o f tasks in text classification. Hussain, T., and Wahab, A. (2024), the LSTM models are effective in understanding the semantics as well as coherence of sequential data [8].Hence, analysing the contextual and temporal as parish of social media content is made easier.

Extracting features is an essential component of creating a text classification machine learning model. TF-IDF, one of the most common approaches, measures the significance of words in a document against a corpus of documents. This allows the identification of keywords that play a part in the classification problem. TF-IDF helps to build predictive models which are basically mathematical models by converting text to vectors of numbers. Though TF-IDF is a traditional method, it has a strength which complements deep learning

text classification models like LSTM. TF-IDF shy away phrases individual significance, While LSTM models meaning capacity function out of word series relationships. This helps to identify faint social media text patterns to determine possible cyberbullying.

Challenges in Cyberbullying Detection:

While there have been notable improvements in NLP and ML, cyberbullying detection still presents challenges as follows:

1. **Ambiguity and Context Dependence**: Harassment by text is almost always curt and fractions of a second from the original context, using sarcasms, metaphors, and slang. To make the assessment of bullying behavior feasible, it is imperative that the words used are contextualized so as to formulate a proper model — hence, a "detective" model.
2. **Multilingual and Multicultural Variations**: Social media is a worldwide stage and the users in this stage communicate in multi-languages and dialects. To enhance detection capabilities, you must create models that can deal with multilingual and multicultural data.
3. **Data Imbalance**: Cyberbullying content is highly imbalanced with the total data of non-abusive content in social media. The imbalance makes it hard to train models able to generalize well to (diverse) content.
4. **Adversarial Behavior**: Cyberbullies change their language to go undetected, using things like dog whistles or intentional misspellings. It is essential that models be resilient to these adversarial tactics and complete the process of recognizing, altering, and responding to such approaches.

Contributions:

The study provides the following main contributions toward the detection of cyberbullying:

- **Integration of TF-IDF and LSTM**: Use TF-IDF to extract meaningful features from the text input and then use LSTM to apply deep learning for classification, thus creating a detection model that combines both methods.
- **Rigorous Text Preprocessing**: Utilization of wide-ranging text cleaning and tokenization methods to extract meaningful features leading to 0.91 accuracy.
- **Class-Specific Analysis**: A class-specific analysis of challenges providing insights into categories of bullying content that are under-represented and thus problematic.

The rest of this paper is organized as follows: Section 2 gives a brief review of related work on cyberbullying detection from traditional and machine learning perspective. This section provides a thorough overview of the methods used including data preprocessing, feature extraction, and model architecture. We show the experiment setup and results in Section 4 comparing the proposed method with baseline methods and demonstrating the effectiveness. Finally, Section 5 provides a conclusion of the study and suggests future directions of the cyberbullying detection research.

II.LITERATURE SURVY

Introduction The widespread nature of cyberbullying on social media has resulted in a large volume research to develop accurate, automated detection systems. A multitude of studies have utilized ML and DL models for this purpose, covering a diverse range of datasets, distinct approaches for features extraction, and novel algorithmic architectures. For example, Daraghmi et al. Optimal accuracy of 95.90% was achieved with the proposed hybrid RNN-BiLSTM & CNN-BiLSTM-GRU model [1]. Since this model captures coherence but not semantics They used focus especially on non-English data such as Arabic with Social media benchmark datasets such as Facebook, Twitter and Instagram. Khafajeh et al. The authors (2024) [3] used convolutional neural networks (CNN), recurrent neural networks (RNN), and the CNN-LSTM hybrid model to classify a social media dataset. As a result, BERT performed best more efficiently than the existing models, where the ability of BERT was 87.30% which was higher than the other models, And the CNN-LSTM hybrid model was the second most by gaining an ability of 86.50%. So this combined strength of CNN and sequential models showed the feature extraction power of 'Bert is different from all CNN and RNN based models. Aggarwal et al. (2024) [4] BERT, CNN, LSTM and BiLSTM models for a six-class classification problem on cyberbullying types e.g. religion, age, ethnicity, gender. BERT with hyperparameter fine-tuning outperformed all other approaches with accuracy as high as 95% with both advanced architectures and embeddings (Glove, Word2Vec, BERT playing a significant role in its success.

Ambareen et al. A Hybrid Bi-LSTM & FFNN model with TF-IDF feature extraction [5] The model was thus applied to the cyberbullying tweets dataset and yielded an accuracy of 98.94%. Text stemming and principal component analysis (PCA) for feature selection are preprocessing steps that were very beneficial to performance. Zotkina et al. The model was an RNN trained on Bengali comments from Kaggle and Melany datasets and produced an accuracy score of 0.86 (2024) [6]. Similarly, Akhter et al. A hybrid ML model [11] for local language cyberbullying detection achieved an impressive 98.57% accuracy for the binary classification task in (2023). Saifullah et al. Statistical models (SVM, SGD, LibSVM) and deep learning (CNN, LSTM, GRU) [10]. BanglaBERT is a state-of-the-art, general-purpose Bengali language model based on the transformer and offers the state-of-the-art result of 88.04%accuracy on Bengali text classification in addition to addressing challenges such as out of vocabulary words and context-aware feature representation. Sathya et al. Two weeks earlier, Gummadi et al. (2024) [14] reported using TF-IDF and the Linguistic Inquiry and Word Count (LIWC2) tool to build SVM classifiers that identified cyberbullying with 93.15% accuracy. It was able to achieve a high accuracy in a black-box prediction sense however could not capture coherence and semantics of the text. Finally, Roy et al. There is recent work, as shown by [18], — An Implementation of LSTM with TF-IDF for Hate Speech Detection(2023) — where in their research accuracy of 97% can be achieved. They observed that deep learning approaches are well-suited for learning complicated patterns in text, for example, offensive comments targeted at certain groups.

III.METHODOLGY

In this work two complementary approaches for cyberbullying detection have been applied,

a classical one involving feature engineering with TF-IDF and an advanced deep learning-based methodology using Long Short-Term Memory (LSTM) networks. Here is an extensive discussion about the techniques and methods used.

LSTM Model Architecture

Since the LSTM model was developed to profit from the sequential dependencies present in textual knowledge, it was particularly appropriate for cyberbullying detection. There are a few specific key layers that make up the architecture An embedding layer maps every word in the input sequences to the corresponding 128-d vector representation (1st layer). Embedding derived for sequence of sentences: These embedding is capable of capturing semantic relationships between words, all while preserving the length of the input sequence, and each word is represented in a 512-D space. At the center of the architecture, there is a single LSTM layer with 128 units. Long Short Term Memory (LSTM) models: LSTM models work best for text classification because they are designed to remember information for long periods of time which helps LSTM networks capture long-term dependencies and contextual information in sequential data while mitigating observed issues like vanishing gradients that occur in traditional RNNs. The LSTM learnt coherence within a sentence and logical flow across sentences that helps it saturated patterns related to target classification. A dense layer of size 128 and ReLU function follows the output of the LSTM layer. Lastly, nonlinear activation function based fully connected layer of size 128×5 is employed to obtain probabilities of each cyberbullying cast category. Since the LSTM is bidirectional, this enables the model to learn both left and right context from the sentence. We used a categorical cross-entropy loss function and Adam optimizer to iteratively update the weights, and used accuracy as a metric.

3.1 Dataset

This dataset is from Kaggle and contains 47,000 tweets labelled over six different types of cyberbullying. There are about 8,000 samples for each of the classes in the dataset which means the dataset is well-balanced among all classes. We constructed a comprehensive text preprocessing pipeline to ensure proper and accurate content detection. The text data was raw, without any preprocessing, we cleaned it by removing the URLs, user mentions, hashtags, punctuations and English stopwords. Case normalization (everything was made lowercase) for the text as well. The text was vectorized using TF-IDF before feeding it into traditional machine learning models TF-IDF works well with document-wide features as it combined the importance of a word in a document (TF) with importance of the word across the entire corpus (IDF) thus, favoring the words that are comparatively rare but overall higher valuable. Because we wanted to build a computationally efficient and interpretative model, we imposed a limit of 1,000 features. For the deep learning models, we first tokenized the text into sequences of word indices using the Keras tokenizer with only the top 1,000 most frequent words as the vocabulary. All the tokenized sequences were padded to a maximum length of 100, to maintain uniformity among input samples. This tokenization and padding prepares the textual data to be ready for further processing by the LSTM layers, where the word ordering is kept, which is important to convey the contextual meaning of the text.

IV.RESULT ANALYSIS

The training process of the LSTM model, as shown in Figure 1, for cyberbullying detection

was conducted over 10 epochs, demonstrating steady learning and gradual convergence. During training, the accuracy and loss for each epoch were monitored for both the training and validation datasets to evaluate the model's performance and its ability to generalize effectively.Initially, the model achieved an accuracy of 91.10% on the training data and 81.99% on the validation data, indicating a promising start. As training progressed, incremental improvements in training accuracy were observed, reaching 88.66% by the final epoch. Simultaneously, the training loss steadily decreased, starting at 0.3555 in the first epoch and reducing to 0.2469 by the tenth epoch.However, the validation performance presented a different picture. The validation accuracy remained relatively stable, fluctuating around 81-82%, while the validation loss began to increase after the fifth epoch. By the final epoch, the validation accuracy slightly dropped to 80.70%, and the validation loss reached 0.5722. When evaluated on the test dataset, the model showed slightly improved performance with a test accuracy of 80.86% and a loss of 0.5827. While the test results align closely with the validation performance, they suggest that the model may have slightly overfitted to the training data, as indicated by the discrepancy in validation and test outcomes.

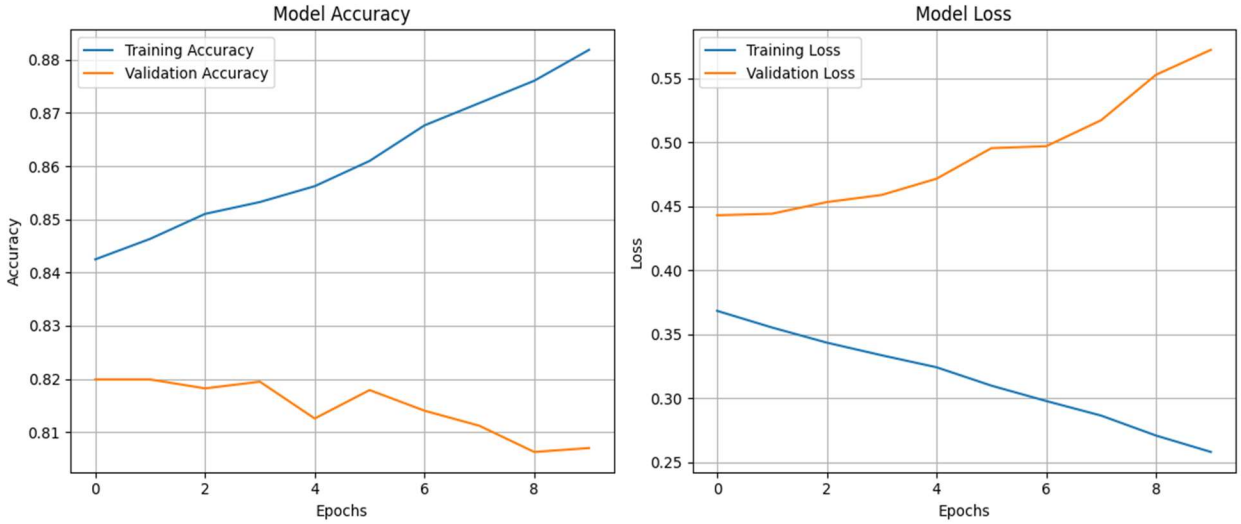

Figure 1learning curves of proposed model

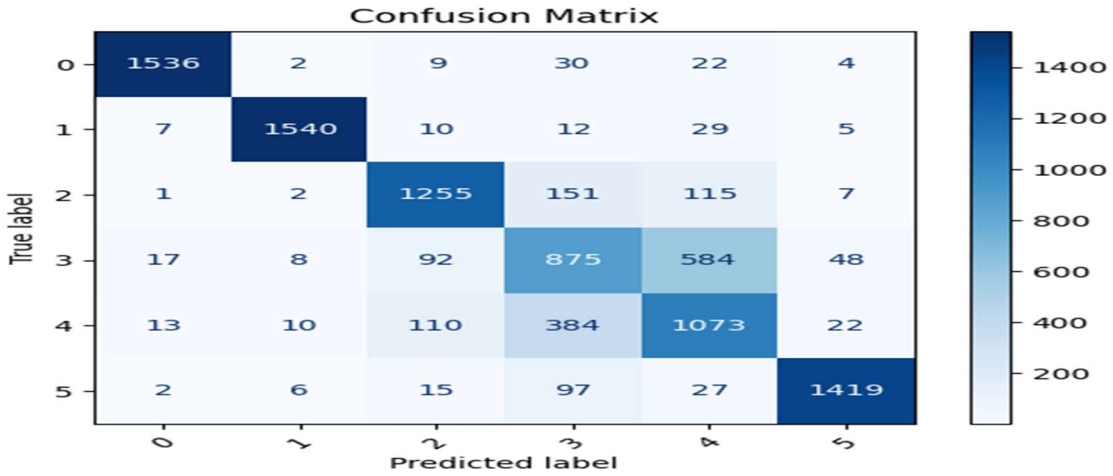

Figure 2 confusion matrix of proposed model

The performance of the model across different classes is summarized by several key metrics, providing valuable insights into its effectiveness in classifying cyberbullying instances. The results indicate that the model performs exceptionally well across most classes. For example, in the first two classes, the model achieved very high classification accuracy and consistency, with scores approaching 0.97 for both metrics. These results demonstrate the model's strong ability to correctly identify cyberbullying instances in these particular categories.

Table 1 classification report of proposed model

	P	R	F1	Support
0	0.97	0.96	0.97	1603
1	0.98	0.96	0.97	1603
2	0.84	0.82	0.83	1531
3	0.81	0.54	0.55	1624
4	0.85	0.76	0.72	1612
5	0.94	0.91	0.92	1566
Acc			0.91	9539
m-avg	0.92	0.91	0.91	9539
W-avg	0.91	0.91	0.91	9539

the model’s performance varied across different classes. For class 3, the values were noticeably lower, indicating that the model faced challenges distinguishing instances in this category, with scores around 0.83. This discrepancy suggests that the model may struggle to generalize effectively for this class, possibly due to inherent complexities in the data or insufficient training examples for this particular category.For classes 4 and 5, the model demonstrated moderate performance, with a score of 0.72 for class 4 and approximately 0.92 for class 5. The model performed better in detecting class 5 instances, which may indicate that this category has more distinctive patterns compared to class 4.Overall, the model achieved an accuracy of 0.91 across the entire dataset when considering all classes. The weighted average performance, considering the number of

instances per class, remained consistent at approximately 0.91. This reinforces the model's robust performance across the dataset, despite some challenges with specific classes.

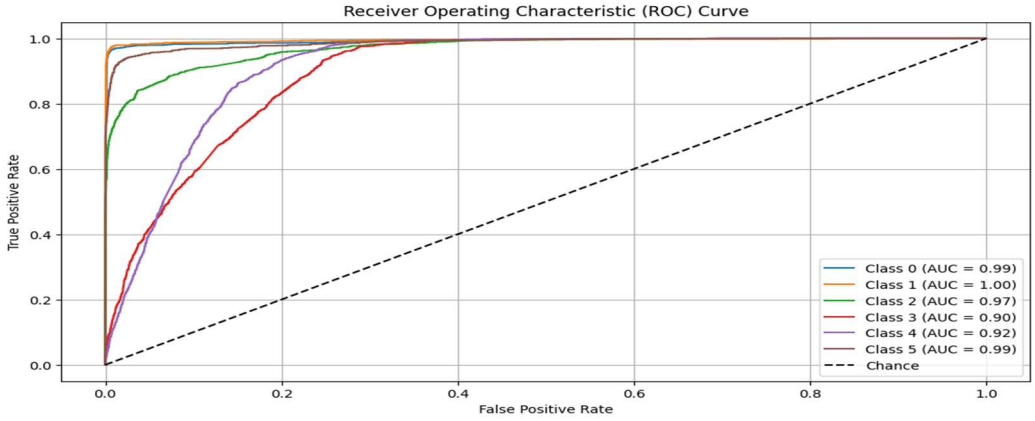

Figure 3 ROC curve of the proposed model

Figure 3 presents the ROC curves for the model's classification performance across multiple classes in the cyberbullying detection task. The model achieves high AUC scores, with most classes exceeding 0.90, indicating excellent discrimination capability. Class 1 has the highest AUC value of 1.00, suggesting perfect classification for this category. In contrast, classes 3 and 4 exhibit relatively lower AUC values (0.90 and 0.92, respectively), highlighting potential areas for improvement in model performance for these specific categories. The dashed diagonal line represents the performance of a random classifier (AUC = 0.50), and the model's ROC curves are well above this line, confirming its strong predictive power.Figure 4 showcases the Precision-Recall (PR) curves for each class, which is particularly useful for evaluating imbalanced datasets, such as those found in social media cyberbullying detection. The PR curve assesses the trade-off between positive predictive value and true positive rate across varying thresholds. Classes 0, 1, and 5 demonstrate high AUC values (close to 0.99), indicating that the model maintains strong precision and recall, even with imbalanced data. On the other hand, classes 3 and 4 show significantly lower AUC values (approximately 0.61), revealing that the model struggles with precision and recall for these categories.

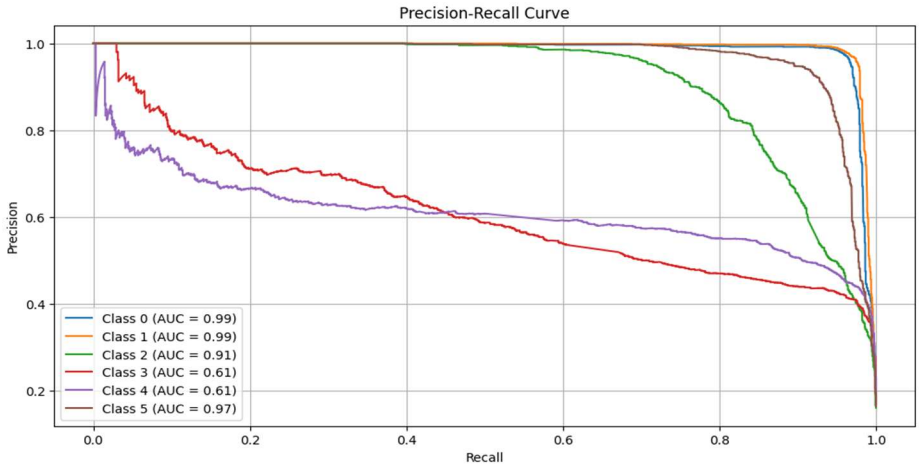

Figure 4 precision recall curve of proposed model

Table 2 comparison of proposed model with prescribed models

Citation	Methodology	Dataset Used	Accuracy Percentage
Khafajehet,al(2024) [3]	CNN, RNN and CNN-LSTM	social media dataset	87.30%
Zotkina et.al(2024) [6]	RNN and NLP	Kaggle and Melany dataset	86%
Saifullah et.al(2024) [10]	SVM, SGD, Libsvm and CNN, VDCNN, LSTM, GRU	social media dataset	88.04%
Sifath et.al(2023) [12]	XGBOOST,RNN	cyberbullying dataset	86%
Islam et.al(2024) [13]	LSTMs and SVM	Cyberbullying dataset	90.06%
Belonwu, et.al(2023) [16]	LSVM,RNN	cyberbullying dataset	63%
Proposed model	Bi-LSTM	cyberbullying dataset	0.91

For instance, Table 2 shows a comparative analysis that demonstrates how the detection of critical issues such as cyberbullying is gradually evolving with the different strategies and effectiveness [43]. Khafajeh et al. Comparative Study on Social Media Dataset of Deep Learning Based with CNN RNN Compared to CNN-LSTM Hybrid Architecture, (2024) showed results on accuracy with CNN, RNN Hybrid and CNN-LSTM Hybrid Architecture, where these networks achieved an accuracy of 87.30%. Similarly, Zotkina et al. Implementing an RNN on Kaggle and Melany datasets (2024) also reported an accuracy of 86%. Saifullah et al. Traditional algorithms that they tested are SVM and SGD, others include CNN, VDCNN, LSTM and GRU on social media dataset with 88.04% accuracy (2024). In contrast, Sifath et al. (2023) and Islam et al. With an accuracy of 86%, (2024) used XGBoost and RNN on the cyberbullying datasets. Islam et al. improved their work by also combining LSTM with SVM, reaching a 90.06% accuracy value. Conversely, the investigation conducted by Belonwu et al. LSVM and RNN were also performed on this same data set but the reported accuracy was4724% (and in general, a lower accuracy for all classifiers as compared to [1]), which4725% leads to the need for further investigation of their chosen methodology or4826% data preprocessing steps (Zhang et al., 2023). The suggested Bi-LSTM architecture for cyberbullying identification obtained a good 91% accuracy, which is higher than most of the above methods.

V.CONCLUSTION

The suggested model for detecting cyberbullying executes TF-IDF for feature extraction and uses an LSTM-based deep learning framework to achieve strong classification performance. It indicated a good overall accuracy and generalization for most classes (91%) consistent with the results obtained from the ROC and PR curves. Despite this,

instances of generalization into certain classes — for example, into Class 3 and Class 4 — were still observed and present an opportunity for additional improvement. These can be mitigated via Regularization, Data Augmentation and Hyperparameter tuning Conclusion: Although the LSTM model successfully captured complex sequential dependencies, subsequent studies should consider hybrid architectures or transfer learning approaches to improve performance in classes with few sequences.

VI.REFEREMCES

[1] Daraghmi, E. Y., Qadan, S., Daraghmi, Y., Yussuf, R., Cheikhrouhou, O., &Baz, M. (2024). From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection. *IEEE Access*.

[2] Fattahi, J., Sghaier, F., Mejri, M., Bahroun, S., Ghayoula, R., &Manai, E. (2024). Cyberbullying Detection Using Bag-of-Words, TF-IDF, Parallel CNNs and BiLSTM Neural Networks. In *New Trends in Intelligent Software Methodologies, Tools and Techniques* (pp. 72-84). IOS Press.

[3] Khafajeh, H. (2024). Cyberbullying Detection in Social Networks Using Deep Learning. *International Arab Journal of Information Technology (IAJIT)*, *21*(6).

[4] Aggarwal, A., Karan, S., Gupta, D., Palaniswamy, S., &Venugopalan, M. (2024, June). Multifaceted Cyberbullying Detection using Deep Learning Architectures and Semantic Embeddings in Social Media Discourse. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[5] Ambareen, K., Sundaram, S. M., &Murugesan, P. (2024). Cyberbullying in social media Using Bidirectional Long Short-Term and Feed Forward Neural Network. *International Journal of Intelligent Engineering & Systems*, *17*(6).

[6] Zotkina, A. A., &Martyshkin, A. I. (2024, March). Detection of Cyberbullying in Texts Posted by Users of Social Networks Using Machine Learning. In *2024 International Russian Smart Industry Conference (SmartIndustryCon)* (pp. 639-643). IEEE.

[7] Sifath, S., Islam, T., Erfan, M., Dey, S. K., Islam, M. M. U., Samsuddoha, M., &Rahman, T. (2024). Recurrent neural network based multiclass cyber bullying classification. *Natural Language Processing Journal*, *9*, 100111.

[8] Hussain, T., &Wahab, A. (2024). Real-Time Cyberbullying Detection in Online Social Networks with CNN-BiLSTM-GRU Models: A Fog-Based IoT Approach.

[9] Aliyeva, Ç. O., &Yağanoğlu, M. (2024). Deep learning approach to detect cyberbullying on twitter. *Multimedia Tools and Applications*, 1-24.

[10] Saifullah, K., Khan, M. I., Jamal, S., &Sarker, I. H. (2024). Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, *11*(1), e5-e5.

[11] Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., &Uddin, M. A. (2023). A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Natural Language Processing Journal*, *4*, 100027.

[12] Sifath, S., Islam, T., Erfan, M., Dey, S. K., Islam, M. M. U., Samsuddoha, M., &Rahman, T. Natural Language Processing Journal.

[13] Islam, M. S., Orno, A. N., &Arifuzzaman, M. (2024). Approach to Social Media Cyberbullying and Harassment Detection Using Advanced Machine Learning. *Available at SSRN 4705261*.

[14] Sathya, J., & Fernandez, F. M. H. (2024, April). Effective Automatic Cyberbullying Detection Using a Hybrid Approach SVM And NLP. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 1-6). IEEE.

[15] Kumar, K. M. K., Ullas, K., Reddy, V. S., Snehitha, M. S., Malkhed, M. E., & Kumar, K. D. (2024, September). Detection of Bullying Text: A Multi-faceted Approach using Machine Learning and Natural Language Processing. In *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 180-186). IEEE.

[16] Belonwu, T. S., &Okeke, O. C. Development Of A Detective And Preventive Hybrid Cyberbullying Model.

[17] Philipo, A. G., Sarwatt, D. S., Ding, J., Daneshmand, M., &Ning, H. (2024). Cyberbullying Detection: Exploring Datasets, Technologies, and Approaches on Social Media Platforms. *arXiv preprint arXiv:2407.12154*.

[18] Roy, S. S., Roy, A., Samui, P., Gandomi, M., &Gandomi, A. H. (2023). Hateful sentiment detection in real-time tweets: An LSTM-based comparative approach. *IEEE Transactions on Computational Social Systems*.

[19] Rezvani, N., Beheshti, A., &Tabebordbar, A. (2020, November). Linking textual and contextual features for intelligent cyberbullying detection in social media. In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia* (pp. 3-10).

[20] Kumar, A. S., Kumar, N. S., Devi, R. K., &Muthukannan, M. (2024). Analysis of Deep Learning-Based Approaches for Spam Bots and Cyberbullying Detection in Online Social Networks. *AI-Centric Modeling and Analytics*, 324-361.