Deep Learning for Multi-Labeled Cyberbully Detection: Enhancing Online Safety

Conference Paper · July 2023

DOI: 10.1109/ICDSNS58469.2023.10245135

6 authors, including:

Rezaul Haque
East West University (Bangladesh)
6 PUBLICATIONS 40 CITATIONS
SEE PROFILE

Piyush Pareek
East West Institute of Technology
91 PUBLICATIONS 237 CITATIONS
SEE PROFILE

Md Babul Islam
Università della Calabria
11 PUBLICATIONS 35 CITATIONS
SEE PROFILE

Mahedi Hassan Ratul
East West University (Bangladesh)
2 PUBLICATIONS 0 CITATIONS
SEE PROFILE

2023 International Conference on Data Science and Network Security (ICDSNS)

Deep Learning for Multi-Labeled Cyberbully Detection: Enhancing Online Safety

1st Naimul Islam
Department of CSE East West University
Dhaka, Bangladesh
naimul.islam.pulak@gmail.com

2nd Rezaul Haque
Department of CSE East West University
Dhaka, Bangladesh
rezaulh603@gmail.com

3rd Piyush Kumar Pareek
Artificial Intelligence and Machine Learning
Nitte Meenakshi Institute of Technology
Bengaluru, India
piyush.kumar@nmit.ac.in

4th Md Babul Islam
School of ICT University of Calabria
Calabria, Italy
mdbabul.islam@dimes.unical.it

5th Imam Hossain Sajeeb
Department of CSE East West Universit
Dhaka, Bangladesh
imamsajeeb25@gmail.com

6th Mahedi Hassan Ratul
Department of CSE
Dhaka, Bangladesh
mahammad.ratul1004@gmail.com

2023 International Conference on Data Science and Network Security (ICDSNS) | 979-8-3503-0159-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDSNS58469.2023.10245135

***Abstract*—Social media platforms offer undeniable benefits, but the preservation of anonymity has led to the emergence of cyberbullying, a concerning social problem. This form of online harassment creates a negative and hostile environment, resulting in decreased user engagement and psychological harm to victims. According to ResearchGate and ScienceDaily, cyberbullying victims in the United States are 1.9 times more likely to commit suicide, highlighting the severity of the issue. However, the current research on cyberbullying detection has been limited to binary/multi-class text classification due to the lack of comprehensive datasets for training and evaluation. To address this gap, we developed a DL-based multi-labeled cyberbully detection system using a dataset of 95,608 social media comments. These comments were categorized into five distinct multi-labeled classes, allowing for a more comprehensive understanding of the different dimensions of cyberbullying. We utilized DL architectures, such as LSTM, BiLSTM, CLSTM, and BiGRU, to develop advanced cyberbully detection systems. By comparing the performance of these DL models with the ML models, we were able to assess the effectiveness and superiority of DL approaches in accurately identifying instances of cyberbullying contents. The CLSTM model, outperformed the others with an exceptional binary accuracy of 87.8% and a macro f1-score of 88.3%. CLSTM's ability to integrate both local and sequential information, coupled with its capacity to capture complex patterns and long-term dependencies, contributes to its superior performance in identifying and classifying cyberbullying instances. By successfully identifying and preventing cyberbullying, our study can contribute to creating a safer and more positive online environment, ultimately enhancing user engagement and satisfaction.**

***Keywords*—cyberbully, text analysis, feature extraction, deep learning**

I. Introduction

Cyberbullying is one of the top ethical issues found on the internet, and can have devastating effects on individuals, causing emotional distress, anxiety, depression, and even leading to self-harm or suicide in extreme cases [1]. According to the organization Enough.Is.Enough, Instagram has the highest incidence of cyberbullying, with over 40% of such incidents taking place on this platform. Facebook and Snapchat are the next most common platforms for harassment, with 39% and 31% of incidents, respectively. The problem is more severe than physical bullying because it is broader and more public, and the victim has nowhere to escape. Cyberbullying happens on social media for several reasons, such as the anonymity provided by social media platforms, a wide audience, and the absence of physical distance. Victims may experience negative emotions like fear, frustration, and embarrassment and feel powerless to stop the cyberbullying. The constant harassment and negative messages can create a toxic environment, making individuals feel unsafe and unable to trust others. Furthermore, cyberbullying can also affect a person's physical health, leading to frequent headaches, insomnia, loss of appetite, and other physical symptoms due to the stress and anxiety caused by the bullying [2].

Recognizing the severity of cyberbullying and intervening to prevent it is critical for safeguarding individuals' mental health and well-being. Early detection can prevent harmful behavior from escalating, provide support to those affected, and promote positive interactions among users, creating a safer online environment. Effective cyberbullying detection can also offer valuable insights into the extent and nature of the problem, enabling researchers and policymakers to identify contributing factors and develop effective prevention and intervention strategies. As cyberbullying can involve various forms of harassing behavior, including multiple sentiments in a single sentence, multi-label classification is the most thorough compared to multi-class classification, as most real-life scenarios have more than one sentiment behind a sentence. The limitations of binary text classification make it inadequate for comprehending the nuances of different cyberbullying classes. Unfortunately, despite significant research [3], [4] on binary/multi-class classification of text in other fields, the absence of a ground truth dataset for training models in the multi-label classification of cyberbullying remains a research gap. To this end, our study has developed a multi-labeled cyberbullying detection system that employs Machine Learning (ML) [13] [14] and Deep Learning (DL) [15] [16] algorithms trained on a dataset of 95,608 abusive social texts categorized into 5 multi-label classes. This research can accurately categorize abusive social media content into multiple labels, providing a more comprehensive approach to cyberbullying detection.

This research contributes to the field by providing a more comprehensive approach to cyberbullying detection, which can help prevent and intervene in cases of cyberbullying and create a safer online environment. Our contributions are as follows:

979-8-3503-0159-5/23/$31.00 ©2023 IEEE

- Collected and built a multi-label cyberbullying dataset of 95,608 posts with ground truth labels: Aggressive, Attacking, Toxic, Sexism, and Acceptable.
- Performed comparative performance analysis to assess the effectiveness and superiority of DL approaches in accurately identifying instances of cyberbullying, with significant label-based classification scores.

II. Related Works

Academics all around the world have been working on developing solutions for the automatic identification of cyberbullying and hate speech, ranging from simple ML models to more complex deep neural network models. Authors in [5] proposed a text-based binary classification of cyberbullying and non-cyber bullying content, using three datasets for their experiment, including two Twitter and Wikipedia datasets. Their proposed method introduces a model combining BiGRU and self-attention mechanism, claiming to address the issue of vanishing and exploding gradient. Their proposed technique outperforms baseline models and achieves an F1- score of 84.9% on the first Twitter dataset, achieves a 96% F1-score for the second Twitter dataset, and achieves a 94.4% F1-score for the Wikipedia dataset. Dadvar et al. [6] claims that a gender-specific approach is better at cyberbullying content classification than the baseline models. They have used My Space posts as their dataset consisting of 381,000 posts from about 16,000 threads where females wrote 34% of posts, and male authors wrote 66%. Binary classification using a support vector machine was done, and incorporating gender-specific features improved the baseline by 39% in precision, 6% in recall, and 15% in F-measure.

A single instance of harassing content can have multiple labels, hierarchically structured based on intensity. It changes regular classification to a more complex task where it's necessary to consider the hierarchical relationships between labels. Pal et al. [7] proposed a multi-label text classification system claiming superior performance over the state-of-the-art models. Multi-Label Text Classification using Attention-based Graph Neural Network (MAGNET) uses five datasets-Reuters 21578, RCV1-V2, AAPD, Slashdot, and Kaggle's toxic comment dataset. MAGNET uses a graph attention network to get the attentive dependencies among labels. They claim better performance over other state-of-the-art models, boasting an F1-accuracy of 89.9%. Rezvani et al. [8] proposes an attention-based context boosted abusive text detection approach. They used an Instagram dataset that had 2188 posts from Instagram along with images, comments, and metadata information for contextual information, and another Twitter dataset containing 7321 Tweets. Upon these texts and contextual information, they have applied a fine-tuned BERT [11] [12] model in a comprehensive and deep architecture that outperforms most state-of-the-art models with 86% accuracy for the Instagram dataset and 85% accuracy for the Twitter dataset. Another approach to abusive tweet detection written in Arabic text is taken by [9], where 8154 tweets were collected using Twitter API, and they chose a lexicon-based approach and a ML-based approach for the classification stage. SVM with resampling methods majorly outperforms the Lexicon-based PMI approach, with 82% accuracy to the PMI's 68%.

After an extensive review of the current literature on cyberbullying and abusive content detection, it becomes apparent that the majority of research has focused on binary and multi-class text classification, neglecting the crucial area of multi-label classification. To the best of our knowledge, no previous studies have specifically focused on developing a multi-label cyberbully detection system. Therefore, this study aims to fill this crucial gap by proposing and evaluating a novel approach for detecting and categorizing abusive content in a multi-label setting.

III. Methodology

A. Dataset Creation and Analysis

The ”Cyberbullying datasets” [10] was collected from various social media platforms such as Kaggle, Twitter, Wikipedia Talk pages, and YouTube, for the purpose of automatic detection of cyberbullying. The collection yielded eight CSV files that were labeled as distinct types of cyber bullying, including hate speech, aggression, sexism, and toxicity. To create a multi-labeled dataset, four binary-labeled datasets (aggression parsed dataset, attack parsed dataset, toxicity parsed dataset, and twitter sexism parsed dataset) were merged, resulting in 95608 texts categorized into five classes: aggressive, attacking, toxicity, sexism, and acceptable. Each text sample was assigned one or two values (0 or 1) indicating the presence or absence of each class. The acceptable class had a value of 1, while 0 was assigned to 73303 texts that did not fall into any of the categories. The remaining 22305 texts were labeled with one or more of the four categories (aggressive, attacking, toxicity, or sexism). This multi-labeled dataset is a valuable resource for training ML models to accurately identify various types of cyberbullying in text. Table I provides examples of the dataset to illustrate how labels were assigned to the text samples.

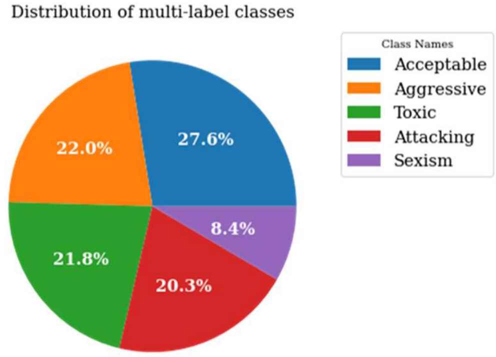

Fig. 1. Multi-label class distribution.

In Figure 1, The dataset's class distribution is presented, indicating the percentage of text samples for each of the five multi-labeled cyberbullying classes. The acceptable class has the largest portion at 27.6%, while sexism has the smallest at 8.4%. The aggressive class has the second-highest percentage at 22%, while the attacking and toxic classes have similar percentages at around 20-21%. The dataset has comparable percentages of data for aggressive, attacking, and toxic classes, except for sexism, which contains a considerably lower amount of data. In order to investigate the interrelationships between the cyberbullying classes, we generated a correlation matrix of the labels, as shown in Figure 2. Our analysis showed negative correlations between the ’Acceptable’ and ’Sexism’ classes with all other cyber bullying classes. ’Aggressive,’ ’Attacking,’ and ’Toxic’ exhibited significant positive correlations with each other, with 'Attacking' and 'Aggressive' having the strongest correlation (90%). The correlation between 'Attacking' and 'Toxic' was 71%, while the correlation between 'Aggressive' and 'Toxic' was 69%.

TABLE I. EXAMPLE OF THE EXPERIMENTAL MULTI-LABELED CYBERBULLY DATASET.

Text	Aggressive	Attacking	Toxic	Sexism	Acceptable
I have a dick, it's bigger than yours! Haha	1	1	1	0	0
just general abuse. Not necessarily gamergate specific.	0	0	0	0	1
I will just push on your belly and a few will come out of your butt!	1	1	0	0	0
Now you are babbling like every Muslim conspiracy idiot.	0	1	0	1	0

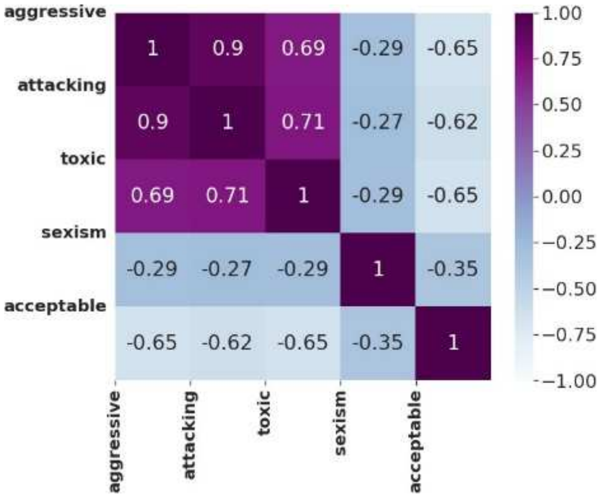

Fig. 2. Correlation of the multi-labeled classes.

B. Data Preprocessing

The data was collected from various online sources, which contain lots of noise and uninformative parts raising high dimensionality problems. We employed various text preprocessing techniques to improve the quality of the text, as outlined below:

1) Basic Operation and Cleaning: The first module of text analysis involves performing basic cleaning operations to prepare the text data for further analysis. These cleaning operations involve removing unimportant elements and transforming certain words to make them more meaningful. To start, the text data is converted into lowercase form using a classical preprocessing technique called lowercasing. This is done to ensure that all words are treated the same way and to avoid duplication of words with different capitalizations. Next, English contractions are expanded to convert the short forms of words into complete forms. After this, URLs, emoticons, mentions, whitespaces, and line breaks are removed from the text data. These elements do not provide significant information and can hinder the accuracy of analysis. The next step involves removing punctuation marks. Although they are frequently used in text, they do not convey any relevant information and their presence can interfere with accurate analysis. Finally, stop words are removed. These are commonly used words that do not carry much information about the role of the text. By removing stop words, the signal-to-noise ratio in unstructured text is increased, which ensures better accuracy and increases the statistical significance of terms.

2) Stemming and Lemmatization: Stemming and lemmatization are two common language modeling techniques that aid in improving information retrieval results. Stemming is a process of removing prefixes and suffixes from words, enabling documents to be represented by the stems of words instead of their original forms. Porter's Stemmer, which comprises five steps that are executed linearly, is a widely used stemming algorithm. However, the application of stemming can sometimes lead to errors in meaning and spelling. In contrast, lemmatization is a technique that reduces a word variant to its root form, using vocabulary and morphological analysis to return words to their dictionary form. Both stemming and lemmatization play a significant role in enhancing the relevance and performance of DL models. Thus, we utilized Wordnet Lemmatizer to obtain the normalized form of words, considering their tagged part of speech, after applying Porter's Stemmer on the sentences.

3) Spell Correction: When it comes to online texts, two types of errors are common, namely semantic and grammatical errors. These errors are typically due to user input errors or a lack of user knowledge. In addition, after performing stemming and lemmatization, some words in the text may have spelling errors. These spelling errors can make the text difficult to read and process, particularly when training a model. Correcting these errors is critical for improving the effectiveness of learning analytics and reducing dimensional ity. To achieve this, we employed Symspellpy, a symmetric spelling correction algorithm package in Python. This algorithm can locate all strings within a maximum edit distance from a large list of strings, which simplifies the complexity of edit candidate generation and dictionary lookup for a given Damerau-Levenshtein distance.

C. Feature Extraction

The process of feature extraction can greatly impact the accuracy and efficiency of text classification, making it a challenging task. ML classifiers require words to be transformed into numerical form, so natural language must be converted into vectors using vector representation. One commonly used technique for feature extraction is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF assigns a weight to each term in a document based on its frequency in the document (term frequency) and its rarity across the entire corpus (inverse document frequency). By considering both the local importance of a term within a document and its global importance across the entire dataset, TF-IDF helps in identifying key terms that distinguish cyberbullying texts from non-cyberbullying ones. The higher the TF-IDF score of a term, the more significant it is in representing the content and context of a document. By leveraging TF-IDF as a feature extraction technique, cyberbullying detection models can effectively capture the distinctive linguistic characteristics and discriminatory terms associated with cyberbullying instances. By utilizing the Keras library, the Word2sequence embedding process can be executed through a pre-built word embedding layer within a neural network. This allows words to be represented in a low-dimensional feature space. To vectorize a text corpus and input it into the neural network, Keras tokenizer API was utilized, requiring three parameters: input length, input dimension, and output dimension. The input length indicates the length of the input sequence for the input layer of the Keras model, and we employed a post sequence padding of 60 lengths for each data instance to ensure uniform sequence length. The input dimension is indicative of the vocabulary size of the text data, which in this work was set to 41322 as there were 41322 unique tokens in the dataset. Finally, the output dimension represents the vector space size of the embedding words, with the size of the output vectors

set to 128 dimensions.

D. Model Development and Classification

Our study involved exploring multi-label classification using several traditional ML and DL algorithms. The ML algorithms we employed included Random Forest (RF), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Multinomial Naive Bayes (MNB). These algorithms were trained with their default parameters to establish a baseline performance for cyberbullying detection. By training these ML algorithms with default parameters, we aimed to determine the initial performance levels and identify the most suitable baseline model for further comparison with DL classifiers.

Opted experimental DL models and their training parameters are described as follows.

1) DL Algorithms: To do this, we randomly divided the dataset into three equal parts, with 90% used for training, 5% for validation, and the remaining 5% for testing purposes. The word embeddings obtained in the previous step were trained separately for each algorithm at the input layer using the training and validation data. LSTM is a type of RNN that uses three gates to control information flow. The input gate decides the importance of data, the forget gate controls transferability of information between network levels, and the output gate determines the next hidden state of the network. BiLSTM uses LSTM units that access both past and future contextual data and contains two parallel layers processing two different textual contexts to capture dependencies. In our study, we employed two BiLSTM layers, where the first layer had 128 internal units and the second had 100 units. The gating mechanism in LSTM and BiLSTM assists in solving the problem of long-term data preservation, which is a challenge in traditional RNNs due to the vanishing gradient problem. The BiGRU model, like BiLSTM, combines information from past and future time steps to predict the current state of a sequence. However, it uses only two gates, update and reset, and trains faster than LSTMs. Our study used two BiGRU layers with different internal units and a C-LSTM model that combines CNN and LSTM. The C-LSTM model extracts maximum features from texts using CNN and uses them as LSTM input. The vectors obtained from word embeddings are passed through a convolution layer with a 3x3 kernel size, 32 filters, and relu activation function and then updated and compressed using max pooling layer.

2) Hyperparameter: The final layer of our proposed multi-label classification models is a dense layer that employs the sigmoid activation function to compute the probability, thereby simplifying the process of obtaining the correct label value. In order to compile all models, we used binaryaccuracy as the metric, rmsprop as the optimizer, with an initial learning rate of 1 10−3, and binarycrossentropy as the loss function. To train our models, we set the number of epochs to 20 and the batch size to 128. To prevent overfitting, early stopping techniques with a patience of 2 epochs and a delta of 0.0001 that monitor the binaryaccuracy of the models were incorporated into the training stage.

E. Evaluation Metrics

Multi-label classification is a complex task as each instance can be associated with multiple labels simultaneously. Evaluation measures for multi-label text classification can be example-based or label-based. In our study, we only considered label-based evaluation measures. These measures assess the predictive performance for each label separately and then average the performance over all labels. Common label-based metrics such as accuracy, recall, precision, and f1 score were obtained using the micro averaging technique, shown in equation 1-4. Where, n is the number of labels, TP_i is the number of true positive instances for label i, FP_i is the number of false positive instances for label i, and FN_i is the number of false negative instances for label.

$$Accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \quad (1)$$

$$Precision = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \quad (2)$$

$$Recall = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \quad (3)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

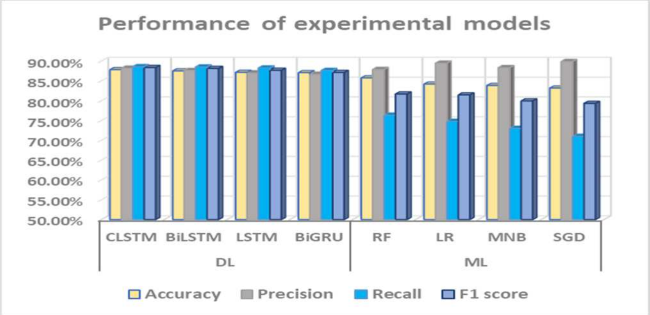

Fig. 3. Performance of experimental models.

Hamming Loss (HL) is a widely used metric for evaluating the performance of models in multi-label classification tasks. HL measures the average number of incorrect predictions made by the model for a given instance. It does so by calculating the ratio of incorrectly predicted labels to the total number of labels, and then averaging this value across all instances in the test set. The HL metric is preferred in multi-label classification problems because it considers the complexity of the task, where an instance can be associated with multiple labels. Equation of HL is shown in equation 5. Where, D is the set of instances, y_i is the ground truth label set for instance i, $\hat{y}_i$ is the predicted label set for instance i, represents the symmetric difference operator, which returns the set of elements that are in either of the sets, but not in their intersection, and D is the number of instances in the set D.

$$HL = (1/|D|) * \sum [i = 1to|D|](y_i \oplus \hat{y}_i) \qquad (5)$$

IV. Results Analysis

A. Performance Comparison

The Table II shows the performance of experimental ML and DL classifiers based on various evaluation metrics. The CLSTM model emerged as the top performer among the all the classifiers, demonstrating superior performance across multiple evaluation metrics. It achieved the highest accuracy (87.83%), precision (88.21%), recall (88.59%), F1 score (88.37%), and the lowest HL (12.30%). The BiLSTM, LSTM, and BiGRU models also yielded comparable results, underscoring the effectiveness of DL algorithms in addressing the task at hand. On the other hand, among the ML models, the RF algorithm demonstrated relatively better performance compared to other ML models. Although the highest-performing ML model, RF, showed promising results with an accuracy of 85.75% and an F1 score of 81.68%, it still fell short in performance when compared to the DL models across most evaluation metrics. Based on the Figure 3, we can see that CLSTM is the best performing model overall, as it achieves the highest scores for all the metrics. Furthermore, our finding suggests that DL techniques are more effective in accurately identifying and classifying abusive content in text, highlighting their superiority over traditional ML approaches in multi-label cyberbully detection.

TABLE II. Performance Of Ml And Dl Classifiers Based On Evaluation Metrics.

Technique	Model	Accuracy	Precision	Recall	F1 score	Hamming
DL	CLSTM	87.83%	88.21%	88.59%	88.37%	12.32%
	BiLSTM	87.51%	87.62%	88.51%	88.13%	12.55%
	LSTM	87.14%	87.02%	88.27%	87.66%	12.86%
	BiGRU	87.06%	86.68%	87.64%	87.11%	13.34%
ML	RF	85.75%	87.88%	76.34%	81.68%	14.31%
	LR	84.19%	89.48%	74.79%	81.45%	14.95%
	MNB	83.84%	88.34%	73.02%	79.93%	15.88%
	SGD	83.17%	89.89%	71.01%	79.31%	16.19%

B. Model Validation

In multi-label text classification, the confusion matrix is a valuable tool for assessing classification accuracy. For each class, a separate confusion matrix was generated in our study. Confusion matrix of the highest performing classifier (CLTSM) is shown in Figure 4. The sexism class had the poorest prediction performance, with approximately 29.2% of the 257 actual sexism classes being misclassified. The model had the best performance on acceptable classes, with only 15.3% being misclassified. The misclassification rates for aggressive, attacking, and toxicity classes were 19.7%, 21.9%, and 15.5%, respectively. Poor training on the low number of cyberbullying texts containing sexism labels resulted in high misclassification rates.

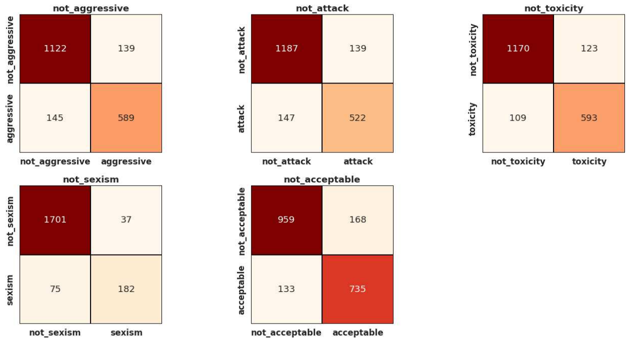

Fig. 4. Confusion matrix of CLSTM model.

In Figure 5, the learning curves for our experimental DL models are depicted, indicating their stability during training. The CLSTM model exhibited the best performance with the highest validation score at the 5th epoch, while BiGRU performed the worst with a final validation accuracy of 86.6%. All models achieved optimized validation accuracy between the 7th and 8th epochs, with a stable rate of change thereafter. However, the loss of LSTM, BiLSTM, and BiGRU was relatively higher than that of CLSTM, indicating a sign of overfitting during long training periods.

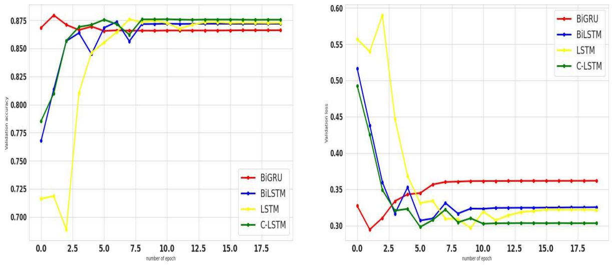

Fig. 5. Learning curve of each DL classifier.

V. Discussion

To accurately identify cyberbullying, we require multi-label text classification that can assign multiple labels to a given text. By accurately identifying and labeling different types of bullying in text, such as verbal abuse, cyberstalking, hate speech, and discrimination, our study contributes to protecting the mental health and well-being of individuals, promoting a safer online environment, and advancing our understanding of this critical issue. The results indicate the CLSTM model had the highest accuracy, precision, and recall scores of 87.8%, 88.2%, and 88.5%, respectively. The results demonstrate that combining various DL algorithms, as we did in our study, can lead to effective systems for multi-label text classification in the cyberbullying detection domain. However, our research can only detect cyberbullying contents from social media text to categorize it into multi-labeled classes, but it fails to identify the bully and take urgent actions. This can only be possible if the research community, law enforcement, and social media platforms work together and share their perspectives and knowledge on the problem. Since there is no publicly available data, the task has become more complex.

VI. Conclusion

Cyberbullying can have serious consequences on an individual's well-being, and detecting it on social media is a complex task. Our research created a multi-labeled dataset for cyberbullying and trained ML and DL models to predict abusive social media content. Previous attempts to detect cyberbullying have mostly used binary or multi-class text classification, which fails to capture its complexity. Our study found that the CLSTM model performed the best out of the four DL classifiers, with the highest accuracy, precision, and recall scores. Furthermore, results suggest that utilizing DL algorithms can result in better accuracy score compared to ML classifiers for developing systems that can classify cyberbullying text into multi-label categories. However, the dataset used in this study lacked the necessary information to perform a comprehensive analysis of user accounts involved in cyberbullying. Moreover, it is imperative to improve the accuracy of the cyberbullying detection system to prevent and intervene in bullying incidents before they occur. Despite these limitations, our study contributes to the current research trend on cyberbullying detection by creating a system that can categorize the intensity of cyberbullying into multi-labeled classes, rather than binary or multi-class categories. Additionally, the dataset used to train our classifiers was collected from various social network platforms, allowing our system to effectively identify cyberbullying patterns across different social networking sites. In future work, we propose to enhance the detection system by incorporating video and image data to build more advanced transfer learning models. Furthermore, we plan to adapt our model to analyze social online behaviors written in different languages (such as Arabic, Bengali, and Hindi) that may negatively impact mental health.

REFERENCE

[1] H. Rosa et al., “Automatic cyberbullying detection: A systematic review,” Comput Human Behav, vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/J.CHB.2018.12.021.

[2] L. Nixon, “Current perspectives: the impact of cyberbullying on adolescent health,” Adolesc Health Med Ther, vol. 5, p. 158, Aug. 2014, doi: 10.2147/AHMT.S36456.

[3] Y. AlHarbi, M. S. AlHarbi, N. J. AlZahrani, M. M. Alsheail, and D. M. Ibrahim, “Using Machine Learning Algorithms for Automatic Cyber Bullying Detection in Arabic Social Media,” Journal of Information Technology Management, vol. 12, no. 2, pp. 123–130, Jun. 2020, doi: 10.22059/JITM.2020.75796.

[4] Almutairi, M. A.-H.-I. J. of Computer, and undefined 2021, “Cyberbullying Detection by Sentiment Analysis of Tweets’ Contents Written in Arabic in Saudi Arabia Society,” koreascience.or.kr, vol. 21, no. 3, p. 112, 2021, doi: 10.22937/IJCSNS.2021.21.3.15.

[5] Y. Fang, S. Yang, B. Zhao, and C. Huang, “Cyberbullying de- tection in social networks using bi-gru with self-attention mecha- nism,” Information (Switzerland), vol. 12, no. 4, pp. 1–18, 2021, doi: 10.3390/info12040171.

[6] M. Dadvar, R. Ordelman, F. De Jong, and D. Trieschnigg, “Improved cyberbullying detection using gender information,” Dutch-Belgian Infor- mation Retrieval Workshop, DIR 2012, pp. 23–26, 2012.

[7] Pal, M. Selvakumar, and M. Sankarasubbu, “Magnet: Multi- label text classification using attention-based graph neural network,” ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence, vol. 2, pp. 494–505, 2020, doi: 10.5220/0008940304940505.

[8] N. Rezvani and A. Beheshti, “Towards attention-based context-boosted cyberbullying detection in social media,” vol. 2, no. 4, pp. 418–433, 2021.

[9] R. Almutairi and M. A. Al-Hagery, “Cyberbullying Detection by Sentiment Analysis of Tweets’ Contents Written in Arabic in Saudi Arabia Society,” Ijcsns, vol. 21, no. 3, pp. 112–119, 2021.

[10] Elsafoury, Fatma (2020), “Cyberbullying datasets”, Mendeley Data, V1, doi: 10.17632/jf4pzyvnpj.1.

[11] Islam, Md Babul, et al. ”Detect deception on banking credit card payment system by machine learning classifiers.” Second International Conference on Cloud Computing and Mechatronic Engineering (I3CME 2022). Vol. 12339. SPIE,2022.

[12] Islam, Md Babul, et al. ”Cost Reduce: Credit Card Fraud Identification Using Machine Learning.” 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2022.

[13] Islam, Md Babul, et al. ”A fiber wireless improved 5G network- based virtual networking system focused on equal bandwidth.” 2021 2nd international symposium on computer engineering and intelligent communications (ISCEIC). IEEE, 2021.

[14] Islam, Md Babul, et al. ”Twitter Opinion Mining on COVID-19 Vaccinations by Machine Learning Presence.” Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022. Singapore: Springer Nature Singapore, 2022.

[15] Islam, Md Babul, et al. ”Pandemic Outbreak Time: Evaluation of Public Tweet Opinion by Machine Learning.” 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET). IEEE,2022.

[16] Islam, Khandaker Sajidul, et al. ”Blockchain Based New E-voting Protocol System without Trusted Tallying Authorities.” 2022 Fifth Inter- national Conference on Computational Intelligence and Communication Technologies (CCICT). IEEE, 2022.

[17] Sayeed, Md Abu, et al. ”Bangladeshi Traffic Sign Recognition and Clas- sification using CNN with Different Kinds of Transfer Learning through a new (BTSRB) Dataset.” 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2023.