Detection of Cyberbullying Using Machine Learning and Deep Learning

Mrs.Deepana M.E.,
Assistant Professor, Department of
Computer Science And Engineering,
Shree Venkateshwara Hi Tech
EngineeringCollege,
Gobichettipalayam.
Email : deepanapostbox@gmail.com

Mr.M.Vigneh,
Student of
Computer Science And Engineering,
Shree Venkateshwara Hi Tech
EngineeringCollege,
Gobichettipalayam.
Email : vigneshmurali215@gmail.com

Mr.R.Palanisamy,
Student of
Computer Science And Engineering,
Shree Venkateshwara Hi Tech
EngineeringCollege,
Gobichettipalayam.
Email : rpalanisamy2002@gmail.com

Mr.I.Solomon francis,
Student of
Computer Science And Engineering,
Shree Venkateshwara Hi Tech
EngineeringCollege,
Gobichettipalayam.
Email:solomonfrancis0304@gmail.com

***Abstract*— As a result of the ease with which the internet and cell phones can be accessed, online social networks (OSN) and social media have seen a significant increase in popularity in recent years. Security and privacy, on the other hand, are the key concerns in online social networks and other social media platforms. In this project proposed system of deep learning approaches for the automated detection and classification 0f cyberbullying on the social media platform , couped with anautomatic blocking mechanism for identified social media accounts.**

***Keywords*— Cyberbullying detection", "Machine learning (ML)", "Deep learning (DL)", "Natural language processing (NLP)", "Social media analysis".**

I. Introduction

Bullying is considered to be an act of abusing a person physically or mentally or verbally while cyberbullying is bullying using digital technologies. Though bullying occurs in specific places such as schools, universities, parks, and workplaces at specific times, cyberbullying occurs at any point of time, anywhere in private online areas. Unlike bullying, cyberbullying doesn't require a large group of people or physical strength for face-to-face interaction. However, it is considered to be a form of harassment that includes pre-teens or adolescents and damages a person.

II. BACKGROUND

A. MACHINE LEARNING

Machine learning (ML) comes under the most booming topic which is artificial intelligence as a branch and also mentions the capability of delivering unmanned or automatic extensive learning which improves the outcomes coming from experiences by detecting the patterns. This technology uses current algorithms as well as datasets in order to develop any computer programs which provide sufficient solutions for the specific problem mentioned and that program will use those algorithms and dataset to learn without any human intervention. The learning process gets started with observing in data given, then identifying the patterns present in the data next creating progress findings by using those algorithms in next coming years based on the identified preexisted patterns. The main aim of using machine learning is that it can makeany electronic device not only computers learn automatically without having any interference from humans or to change results correspondingly. Machine learning algorithms can analyze huge amounts of data which results in high accuracy in a small amount of time.

B. DEEP LEARNING

Deep Learning is one of the main techniques which is used in machine learning. In deep learning, data models are designed in such a way that they bind to the particular task. Deep Learning has applications in various fields including classification and recognition of images, recognition of patterns also in the field of making decisions. These algorithms requires large dataset to achieve better accuracy.

III. RELATED WORKS

Cyberbullying must be detected not only to avoid adverse physical but also effects. Many researchers are working continuously to develop a model usingefficient cyberbullying detection techniques. This section primarily includes researchrelated works done in the field of cyberbullying.

Md Manowarul Islam et al. [1] developed a model, by employing Naive Bayes, Support vector machine (SVM), Decision Tree, and Random Forest on two distinct datasets to detect cyberbullying. SupportVector Machine provided higher performance whenthey used TF-IDF as a feature extractor.

Md Manowarul Islam et al. [1] developed a model, by employing Naive Bayes, Support vector machine (SVM), Decision Tree, and Random Forest on two distinct datasets to detect cyberbullying. One from Twitter, while the other from Facebook comments and posts.Authors were able to get better results for both Facebook and Twitter datasets. Support Vector Machine provided higher performance when they used TF-IDF as a feature extractor.

Haidar et. al [2] presented an study for English and Arabic languages to detect cyberbullying by collecting texts from Facebook and Twitter platforms. They used Support Vector Machine and Naive Bayes classifiers to examine the collected datasets. Around 90.1% Precision was got from Naive bayes and 93.4 %precision for SVM. Baliram Chavan et al. [3] likewise designed a model for Machine Learning to identify and also to check cyberbullying on Twitter platform that utilizes Naive bayes and Support vector machine classifiers. They accumulated the data using Twitter and got accuracy of 71.25 %.

Many more studies on identifying cyberbullying exist, such as [4], in which G. A. Leon-Paredeset et. al, used machine learning algorithm by collecting spanish texts from Twitter and got an accuracy of 93%. Work done by Ali et al.[5] achieved 80% accuracy by using Machine learning algorithms on the three datasets collected. If deep learning techniques had been utilised in these studies, the accuracy and text classification would have improved.

Varun Jain et al. created a model for detecting cyberbullying using a big dataset in [6]. The researchers designed and evaluated the system using a binary classification problem, in which they recognised two categories of cyberbullying: On Twitter for hate speech and on Wikipedia for personal attacks and classified the content as cyberbully or not. They discovered that employing Natural Language Processing (NLP) approaches and procedures resulted in 90% accuracy using simple ML algorithms for the Hate speech dataset. As tweets with Hate speech comments or posts consist of bad language which turned out to be easily detectable.

Rounak Ghosh et al. [7] developed a model for Cyberbullying detection in Indian Language. The Authors wanted to build a system that detects Cyberbullying in the Bengali language which is considered to be outspoken. In the data preprocessing step, they used a stop word filter and transformed all the data into lower case and further tokenization was carried out. To extract the feature from the input text comments, they used TF-IDF method. In Final stage, ML algorithms were used for classification such as, Passive Aggressive Classifier, Support Vector Machine (SVM), Random Forest Logistic Regression. After the classification, the Passive-Aggressive algorithm got high accuracy for N- Gram level features. For the Support Vector Machine algorithm for the word level feature extraction method, they were able to get a better result.

Furthermore, Vijay Banerjee et al. [8] used deep learning methods to construct a model for detecting cyberbullying. For tweet classification, the authors proposed using Word vectors by feeding them into Convolutional neural network (CNN). The authors collected dataset from multiple online social media websites to validate their findings. They implemented their project in python and TensorFlow. For the neural network model in this research, it was implemented using Keras which is a library that works on top of TensorFlow. They were able to achieve 93.97 % accuracy upon testing their model.

Work done by A. M. Syed et al. [9] designed a new technique using deep learning to identify cyberbullying. They collected 39000 tweets using the Twitter API and achieved a maximum accuracy of 95%. In their study, they exclusively used tweets. In a similar line, Srivastava et al. [10] investigated the usefulness and deep learning performance to identify cyberbullying. For the data collected from Kaggle, bidirectional LSTM surpassed the other four deep learning models with an accuracy of 82.18%. Although the accuracy achieved by them is adequate, but we outperform them in terms of accuracy.

Comparative literature review for sentiment analysis conducted Jain et al. [11] proved, many researchers were able to achieve 90% accuracy for machine learning algorithms. With Deep Learning and Combination classifiers one could achieve 93.1-94.9%accuracy. This research inspired us to compare and experiment the performances of algorithms in Machine learning and Deep learning.

Work by April Kontostathis et al. [12] used Machine learning algorithms by collecting data from Formspring. me platform to detect cyberbullying. This website is used to ask questions and answer them. To label the truth data sets, they used Amazon’s Mechanical Turk service. Data is divided into two categories that is ”yes” or ”no”. Two separate training sets were recovered, one for counting data and the other for normalising data. For training sets, the J48, JRIP, IBK, AND SMO ALGORITHMS were used. A decision tree is created using J48. Surprisingly, the overall accuracy rate was 81.7 %. But they used around 2600 text input for their study which is very less but in this study more than lakh comments were used as an input to achieve better accuracy.

Andrew M. Dal et al. [13] used CNN and LSTM to create a sequence learning supervised model. They discussed SoftMax combination in multicategory classification using both oneversus-all and one-versus-one classifiers. K. Duan et al. [14] study demonstrates how to apply the binary classification approach to multi-category classification efficiently. Their research demonstrates the application of the binary classification approach in classifying multiple categories. Quanzhi Li et al. [15] designed a classification strategy for sentiments in tweets on schemes like weighting, the negation of texts.

A survey [22] showed, Instagram ranked highest in cyberbullying activities. Around 42% of people who participated in the survey experienced harassment on this social media platform.

The works discussed above are all excellent, yet they are all imprecise. The dataset, which contains 184397 English texts, is one of the study’s distinctive features. This study not only meant to apply different algorithms but also comparing them to detect cyberbullying in English texts. We go over our recommended methodology in-depth in Section IV.

I. PROPOSED METHODOLOGY

A. Dataset

The present study uses a dataset from the Mendeley data website consisting of 159,686 comments, out of which 144,324 were labelled non-bullying and 15,362 as bullying. The presence of imbalanced data made the classifier detect the cyberbullied comments with low accuracy. For this reason, 24,708 bullied comments were added to the considered dataset. A sample labelled text is shown in Table 1.

TABLE I
SAMPLE INPUT TEXT DATA

Texts	Label
I hate to admit that you are a piece of shit	1 (Bullying)
Please shut the fuck up!! you asshole	1 (Bullying)
Guys lets go on a vacation for a week	0 (Not Bullying)
Can you please look over my cats for a day?	0 (Not Bullying)
You go and fuck your dad	1 (Bullying)
Hello Beautiful	0 (Not Bullying)

B. Work flow

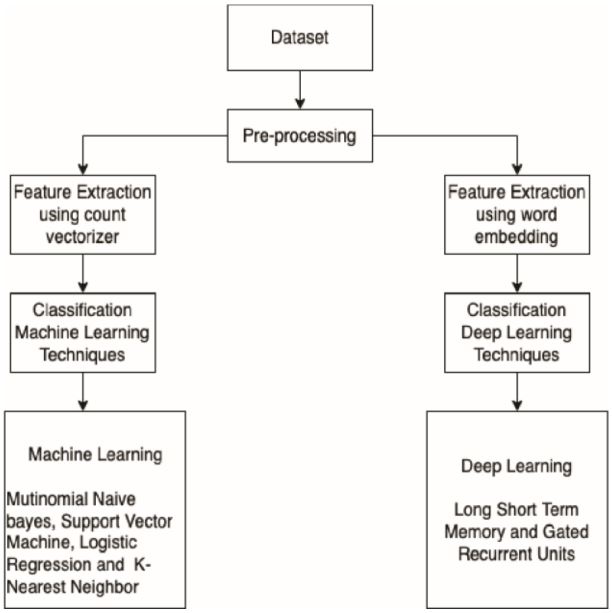

Dataset
Pre-processing
Feature Extraction using count vectorizer
Feature Extraction using word embedding
Classification Machine Learning Techniques
Classification Deep Learning Techniques
Machine Learning
Mutinomial Naive bayes, Support Vector Machine, Logistic Regression and K-Nearest Neighbor
Deep Learning
Long Short Term Memory and Gated Recurrent Units

Fig. 1. Flow diagram of present study

The steps followed in the study are shown in Fig. 1. Before building the classification models using both machine learning and deep learning techniques, a pre-processing task has been carried out for the better performance of the model. In this process, count vectorizer is used in the case of machine learning and word embedding is used in the case of deep learning technique. To maintain uniformity in the dataset, all the sentences are converted from title case or capital case into the lower case as part of pre-processing. In addition to that tokenization is carried out to generate tokens from the text which could make the model understand the context. Finally, stopwords and punctuation were removed from the text which is considered to be an important task in pre-processing the reason that these things don't contribute to the process of developing a model.

Since the machine learning techniques require numeric input, each sentence in the data set is converted into a vector form using count vectorizer which converts based on the frequency of each word. For deep learning techniques, word embedding is used which follows a similar representation for the words with similar meanings. After extracting and importing features using these approaches the processed data are sent to both machine learning and deep learning techniques.

In both the approaches 80% of the data is used for training and 20% is used for validating the model.

C. Machine Learning Algorithms

Four Machine learning based classifiers are built using preprocessed data namely K-Nearest Neighbor, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine.

1) Multinomial Naive Bayes: This classifier works on the Bayes theorem which is represented using Equation 1. Since the dataset consists of text sentences, the probability of each class label is computed and returns the class label which has the highest probability.

$$P(h|b) = P(h|a) * P(h)/P(b) \quad (1)$$

2) Support Vector Machines (SVM): This is one of the powerful classification techniques that come under supervised learning which can work better even if the number of dimensions is greater than the number of samples. This technique is considered to be a memory-efficient technique which determines the best decision boundary between the two classes with the help of a support vector. To categorise data points in Ndimensional space, SVM finds a hyperplane where the number of features is denoted by N. Among the several alternatives, SVM selects the hyperplane with the biggest margin. Increasing the margin distance allows for more exact classification of the following data points.

3) Logistic Regression: For two-class classification, logistic regression is a popular classification method. The logistic sigmoid function is used in logistic regression to transform any actual value into a number between 0 and 1. It is used to translate predicted values to probability. Equation 2 is for sigmoid function.

$$S(Z) = \frac{1}{1+e^{-}Z} \quad (2)$$

S(z) can be represented by the output range of 0 to 1. The function's input is z, and the natural log's base is e. In order to translate the value returned by the function into a discrete class, a threshold value is defined above which the values will be classified as class 1 and below which the values will be classified as class 2. The mapping is shown in Equations 3 and 4.

$$P \geq 0.5; class = 1 \quad (3)$$

$$P > 0.5; class = 0 \quad (4)$$

4) K-nearest Neighbor(KNN): This algorithm is an instance-based learning technique for multi-class problems and uses the metric distance between a fresh sample and its neighbour to classify it. From the training set, determine the K-nearest neighbours and assign an item to the class that is most common among its nearest neighbours denoted by k. This classifier is a non-parametric lazy learning algorithm that makes no assumptions about the distribution of the underlying data.

D. Deep Learning Algorithms

Deep learning is the subset of machine learning techniques which are used to imitate the human brain. Unlike machine learning algorithms these algorithms use numerous layers to build the model. Deep learning algorithms require a larger dataset to get good accuracy. Such two popular deep learning techniques for text classification are namely LSTM and GRU, used in the present study.

1) One of the popular deep learning neural networks is Recurrent Neural Networks also known as RNN has certain drawbacks such as vanishing gradient problems and short-term memory due to which the classifier may not perform better in case of longer sentences. LSTM is a type of RNN which performs fairly better due to its long term memory. It can have multiple hidden layers and pass on the relevant information through every layer and discards unwanted information. It keeps track of dependencies across long gaps [18] and prevents gradients from disappearing. The forget gate is the LSTM's middle layer, and it determines which data should be normalised and which should be forgotten. An input gate modulates the inputs of each memory cell, whereas an output gate modulates the output. The architecture of LSTM is shown in Fig. 2. The present study uses 32 hidden layers with the rectified linear unit as an activation function and a sigmoid activation function in the output layer.

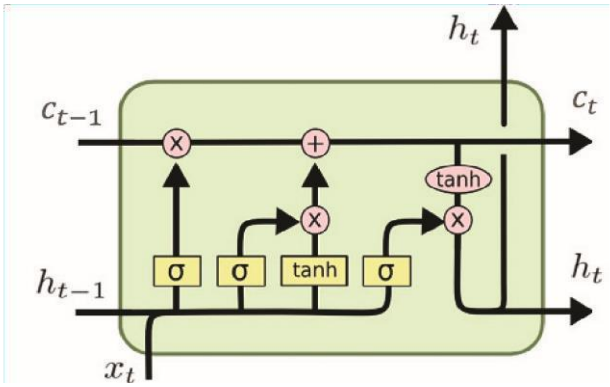


Fig. 2. LSTM architecture

2) Another type of RNN architecture called Gated Recurrent unit also known as GRU uses two different gates namely reset and update. To combine the new input along with the previous memory reset gate is used and how much of the previous memory to be retained is handled by the update gate. Though GRU is similar to LSTM, it trains faster and performs better. It also solves the vanishing gradient problem. The architecture of GRU is shown in Fig. 3.

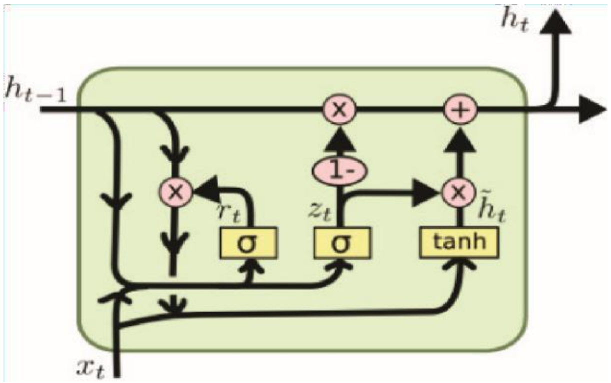
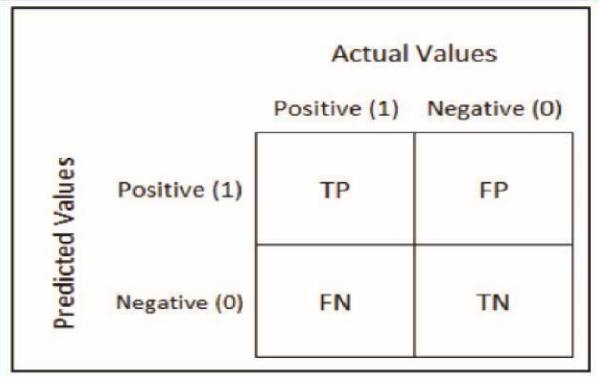


Fig. 3. GRU architecture

II. Results and Discussion

To test the model's performance in terms of conceptual soundness, a model validation has been carried out using test data. The models which are built using training data are evaluated using metrics including precision, recall, accuracy and f1-score and the accuracy details are recorded in Table III while other metrics details are recorded in Table IV and V. The information required to compute such metrics are presented in Table II and the corresponding confusion matrix is shown in Fig. 4 in terms of True Positive(TP), False Positive(FP),False Negative(FN), and True Negative(TN). The values of TP, TN, FP, FN for the test data are recorded in Table II. The corresponding graph is shown in Fig. 5 for the Machine learning approach and shown in Fig. 6 for deep learning approach.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 4. Confusion Matrix

Equations 5, 6, 7, and 8 are used to calculate the accuracy, precision, recall, and f1-score.

$$Accuracy = \frac{TP + FN}{TP + TN + FN + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

TABLE II
CONFUSION MATRIX RESULTS OF CLASSIFIERS

	Algorithm	TP	FP	FN	TN
Machine learning	Support Vector Machines	27,924	942	982	7,032
	Logistic Regression	28,274	592	1,446	6,568
	Multinomial Naive Bayes	28,354	512	1,580	6,434
	K-Nearest Neighbor	26,978	1,888	2,683	5,331
Deep Learning	Gated Recurrent Units	28,204	820	847	7,009
	Long short-Term memory	28,050	816	921	7,093

As shown in Table II, Gated Recurrent Units exhibited the best performance by identifying 28,204 positive labelled correctly on testing data and 7,009 test data were labelled as negative. The lowest result was got when K-nearest neighbour is applied. As it correctly classified 26,978 positive and 5,331 negative labelled testing data.

TABLE III
ACCURACY RESULTS OF MACHINE LEARNING AND DEEP LEARNING CLASSIFIERS

	Classifier	Accuracy
Machine Learning	Support Vector Machines	94.78%
	Logistic Regression	94.47%
	Multinomial Naive Bayes	94.32%
	K-Nearest Neighbor	87.60%
Deep Learning	Gated Recurrent Units	95.47%
	Long Short-Term memory	95.29%

From the Table III we can observe, one of the Deep learning classifiers Gated Recurrent Unit performed well with an accuracy of 95.47% and Support Vector Machine got highest accuracy of 94.78% among machine learning classifiers applied.

TABLE IV
PRECISON, RECALL AND F1-SCORE RESULTS FOR MACHINE LEARNING CLASSIFIERS

	Classifier	NB	B
Precision	Support Vector Machines	0.97	0.88
	Logistic Regression	0.95	0.92
	Multinomial Naive Bayes	0.95	0.93
	K-Nearest Neighbor	0.91	0.74
Recall	Support Vector Machines	0.97	0.88
	Logistic Regression	0.98	0.82
	Multinomial Naive Bayes	0.98	0.80
	K-Nearest Neighbor	0.93	0.67
F1-Score	Support Vector Machines	0.97	0.88
	Logistic Regression	0.97	0.87
	Multinomial Naive Bayes	0.96	0.86
	K-Nearest Neighbor	0.92	0.70

TABLE V
PRECISON, RECALL AND F1-SCORE RESULTS FOR DEEP LEARNING CLASSIFIERS

	Classifier	NB	B
Precision	Gated Recurrent Units	0.97	0.90
	Long Short-Term memory	0.97	0.90
Recall	Gated Recurrent Units	0.97	0.89
	Long Short-Term memory	0.97	0.89
F1-Score	Gated Recurrent Units	0.97	0.89
	Long Short-Term memory	0.97	0.89

Table IV and V displays the metrics of six classifiers,

Precision, recall, and f1-score percentage for Machine learning and Deep learning algorithms applied in the study respectively. In Table V, 'B' represents Bullied category and 'NB' respresents non bullied category.

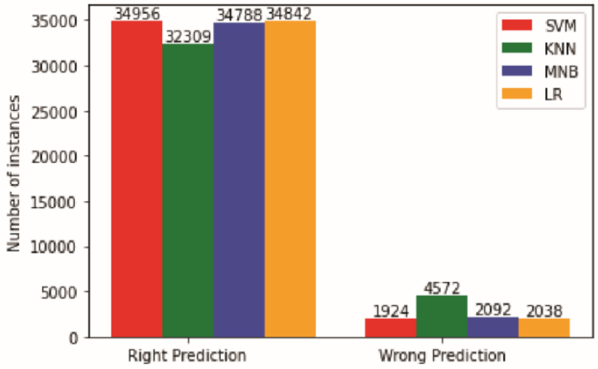


Fig. 5. No. of right and wrong predictions in Machine Learning Algorithms

As shown in Fig. 5 among four Machine Learning algorithms applied, Support Vector Machines performed well by correctly classifying 34,956 occurences and incorrectly identifying 1,924 cases. And in Fig. 6 Gated Recurrent Units (GRU) performed well by correctly classifying 35,213 occurrences and incorrectly identifying 1,687 cases.

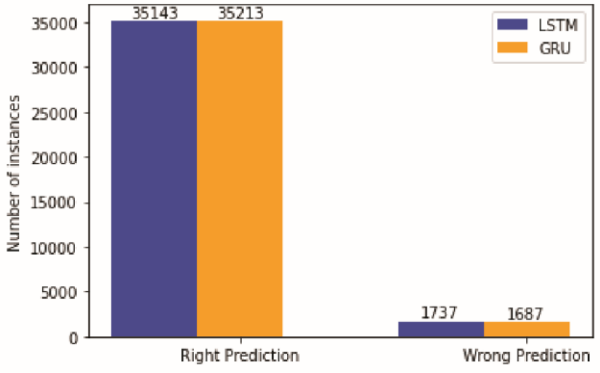
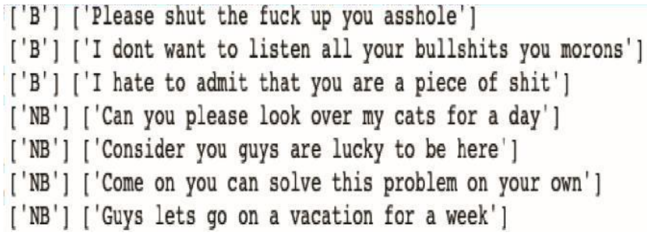


Fig. 6. No. of right and wrong predictions in Deep Learning Algorithms

The sample output given by the Multinomial Naive Bayes model for the new set of test comments shown in Fig. 7 where ’B’ represents bullied comments and ’NB’ represents nonbullied comments.



```
['B'] ['Please shut the fuck up you asshole']
['B'] ['I dont want to listen all your bullshits you morons']
['B'] ['I hate to admit that you are a piece of shit']
['NB'] ['Can you please look over my cats for a day']
['NB'] ['Consider you guys are lucky to be here']
['NB'] ['Come on you can solve this problem on your own']
['NB'] ['Guys lets go on a vacation for a week']
```

Fig. 7. Sample output for new test comments

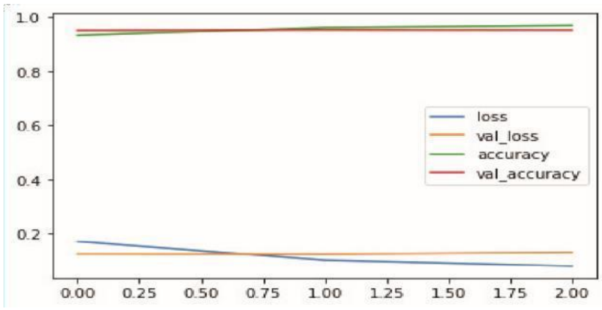


Fig. 8. Accuracy and loss graph of LSTM

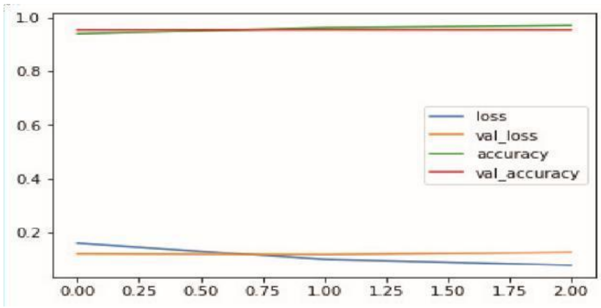


Fig. 9. Accuracy and loss graph GRU

Fig. 8 explains the loss, validation loss, accuracy and validation accuracy for the LSTM model and the same for GRU is shown in Fig. 9.

III. Conclusion and Future Scope

It is observed that SVM is performing better in the case of the machine learning approach and GRU is slightly performing better compare to LSTM. However, it is also clear that deep learning techniques are outperforming compare to machine learning techniques.

It is found that among all the techniques that are applied in the present study, Gated Recurrent Units is performing better with an accuracy of 95.47%.

The present study considers cyberbullying and non-cyber bullying as two different categories further we can also explore various forms of cyberbullying as future work.

Acknowledgment

We are very grateful to Dr.T.SENTHIL PRAKASH MEPhD, Professor and Head Department of Computer Science and Engineering , for the aspiring suggestion, invaluably constructive criticism and friendly advice.

We wish to express our gratefulness to our guide Mrs.D DEEPANA ME., Assistant Professor of Computer Science and Engineering, for his invaluable guidance and constructive suggestions

REFERENCES

[1] Md Manowarul Islam, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin Uzzal Kumar Acharjee., 2020, Cyberbullying Detection on Social Media Network using Machine Learning Approaches. In 2020 Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) IEEE.

[2] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," Adv. Sci. Technol. Eng. Syst. J., vol. 2, no. 6, pp. 275–284, 2017.

[3] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, "Detecting A twitter cyberbullying using machine learning," in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020.

[4] G. A. Leon-Paredeset al., "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language," in 2019 IEEE CHILEANConference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2019.

[5] A. Ali and A. M. Syed, "Cyberbullying Detection using Machine Learning," Pakistan Journal of Engineering and Technology, vol. 3. 2, pp. 45–50, 2020.

[6] Varun Jain, Vishant Kumar, Vivek Pal, Dinesh Kumar Vishwakarma., 2021, Detection of Cyberbullying on Social Media Using Machine Learning. In 2021 Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE.

[7] T Senthil Prakash, V CP, RB Dhumale, A Kiran., "Auto-metric graph neural network for paddy leaf disease classification" - Archives of Phytopathology and Plant Protection, 2023.

[8] T Senthil Prakash, G Kannan, S Prabhakaran., "Deep convolutional spiking neural network fostered automatic detection and classification of breast cancer from mammography images" - Research on Biomedical Engineering,

[9] TS Prakash, SP Patnayakuni, S Shibu., "Municipal Solid Waste Prediction using Tree Hierarchical Deep Convolutional Neural Network Optimized with Balancing Composite Motion Optimization Algorithm" - Journal of Experimental & Theoretical Artificial …, 2023

[10] TS Prakash, AS Kumar, CRB Durai, S Ashok., "Enhanced Elman spike Neural network optimized with flamingo search optimization algorithm espoused lung cancer classification from CT images" - Biomedical Signal Processing and Control, 2023