Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

STAGED RECURRENT NEURAL NETWORK AND LONG SHORT TERM MEMORY BASED DEEP LEARNING ARCHITECTURE TOWARDS CYBERBULLYING DETECTION AND SEVERITY CLASSIFICATION IN ONLINE SOCIAL MEDIA PLATFORMS

Dr. S. UMA[1], Dr. C. THILAGAVATHY[2], S. GOKILAVANI[3], M. VINU[4], S. PRIYADHARASHINI[5] Dr.ARCHANA NANDIBEWOOR[6], M. SHANTHINI[7], G. RAJASEKAR[8]

[1]Professor, Department of Computer Science, Dr.N.G.P Arts and Science College, India

[2]Assistant Professor, Department of Information Technology, CMS College of Science and Commerce (Autonomous), India

[3]Assistant Professor, Department of Computer Science Engineering Rathinam Technical Campus, India

[4]Research scholar, Department of electronics and communication engineering, Velammal College of Engineering and Technology, India

[5]Assistant Professor, Department of Computer Science, Dr.N.G.P Arts and Science College, India.

[6] Associate Professor & HOD, Department of Computer Science and Business System, Dayananda Sagar College of Engineering, Bangalore

[7]Senior AI specialist, Dagslore Technology Solutions, India

[8]Chief Executive Officer, Dagslore Technology Solutions, India

E-mail: [1]uma.s@drngpasc.ac.in, [2]thilaga.c@gmail.com, [3]gokicse27@gmail.com, [4]vinusreee@outlook.com, [5]priyadharshini.s@drngpasc.ac.in, [6]narchana2006@gmail.com, [7]shanthini@dagslore.com, [8]rajasekar@dagslore.com

ABSTRACT

Online social media platforms gaining accelerated popularity in the last decade towards experiencing their views and opinion about personnel, product and social information. Cyberbullying is becoming severe concern to the online social media as it affects multiple categories of peoples and its societies. In literatures, many automated methods have been employed by researchers to detect the cyberbullying. Those architectures examine the social and cultural phenomena to identify the behavioral pattern of the post. Despite of the several advantages, those model fails considers contextual factors and computational complexities which leads to cold start issues, uncertainty, class imbalance problem, changes in velocity and volume. To mitigate those challenges, a new architecture has to be established to alleviate it by detecting and classifying the severity of the cyberbullying post. Stacked Recurrent Neural Network with Long Short Term Memory is designed and developed to process temporal and time varying data on computing the long term dependencies among the various users. Particular model was trained to extract the contextual, temporal, spatial and sentimental features of data. Those extracted features are represented as feature vector and it is processed further to construct the feature mapping using long short term memory layer of the model. Subsequently, the generated feature map is processed with classification function on the connected layer to detect and classify the severity of the post as mild, moderate and strong. Experimental analysis of the model is performed using twitter dataset in the python environment and it exhibits better processing capabilities and computational capabilities on reducing cold start and class imbalance issues.. Further performance analysis of the model is performed to compute prediction and classification accuracy which produces improved classification accuracy of 98.9 percent and predictive accuracy of 98.7 against different user characteristics. Finally it proved to be outperforming compared to conventional approaches

Keywords: *Cyberbullying, Recurrent Neural Network, Long Short Term Memory, Online Social Media, Twitter dataset*

1. INTRODUCTION

Rapid advancement of the online social media has provided a provision to its user to post and respond to the user generated content in form of views and opinions. Those advancements cause many detrimental consequences in form of cyberbullying and privacy risk to their personnel or social information's. Cyberbullying is a severe concern which initiates towards harassing, threatening, humiliating, intimidating, manipulating and controlling victims [1]. Those concerns will severely impacts victim's mental health in form of depression, negative emotions and suicidal thoughts. Thus it becomes mandatory to detect and classify the opinions posted to the user generated contents[2].

Traditionally many researchers have initiated designing and deploying multiple model using artificial intelligence and natural language processing techniques on strongly examining social and cultural phenomena of the user to detect and classify cyberbullying post towards effective mitigation. Despite of the several advantages those architectures fail to consider informal nature of ambiguous expression of intricate languages as it makes it complex to process[3]. Especially it leads to cold start, curse of dimensionality and class imbalance challenges due to changes in velocity and volume of the cyberbullying post. In this paper, a new deep learning architecture entitled as "Stacked Recurrent Neural Network with Long Short Term Memory" is designed and implemented to alleviate cyberbullying post to user generated content by detecting and classifying it as mild, moderate and severe.

Initially preprocessing step is performed using stop word removal, stemming, tokenization and token weighting. Weighted Token is projected to Long Short Term Memory model which store state information of the data as input gate. Next, word vector is propagated to Stacked Recurrent Neural Network which is composed of deep convolution layer to extract the contextual, temporal, spatial and sentimental features of weighted tokens and hidden layer to generate the dependency map of the feature. Dependency map is stored as state information in the LSTM model which performance concatenation and averaging through various gating mechanism.

Finally fully connected layer of the model uses the activation function to linearize the dependency map from output gate of the LSTM and softmax function uses those gate information's to detect and classify the cyberbullying post into different severity classes. Rest of the article is organized as follows; section 2 illustrates the related works of the cyberbullying detection and classification using artificial intelligence approaches along its experimental and performance analysis. Section 3 defines the architecture of the proposed deep learning approach entitled as Stacked Recurrent Neural Network with Long Short Term Memory Model towards detection and classification of the cyberbullying post to the user generated content in the online social networks along mitigating the challenges of the existing architectures. Section 4 represents the experimental and performance analysis of the proposed architecture against various conventional methods is performed using twitter dataset along incorporating confusion matrix and its performance measures. Finally section 5 concludes the article with major findings.

2. RELATED WORK

In this section, related artificial intelligence architectures towards cyberbullying detection and classification were analyzed against various architecture elements and along its experimental and performance analysis. Those architectures analysis is as follows.

2.1 Optimization Recurrent Neural Network using Dolphin Echolocation Algorithm towards Cyberbullying detection and Classification

In this literature, optimized recurrent neural network using dolphin echolocation algorithm is employed towards detecting and classifying cyberbullying post to user generated content as harassment, threatening and humiliating[4]. Cyberbullying post towards model training and model testing is extracted from twitter dataset. Initially, data preprocessing and data annotation is carried out to the collected data. It is performed to handle dynamic nature of the tweet in form of short text. Next model training is performed towards detection and classification along inclusion of feature extraction and feature selection mechanism. Especially Hidden layer of the model performs both feature extraction process through kernel function and feature classification process through softmax function along optimizing it layer through dolphin echolocation algorithm.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Further experimental and performance efficiency of model is computed using cross fold validation. Finally specified model provides 92.8% accuracy.

2.2 Transformer based model towards Cyberbullying detection and Classification

In this literature, transformed based model represented as Optimized BERT is employed towards detecting and classifying cyberbullying post to user generated content as harassment, threatening and humiliating. Cyberbullying post towards model training and model testing is extracted from twitter dataset. Initially, data preprocessing and data annotation is carried out to the collected data. It is performed to handle dynamic nature of the tweet in form of short text. Next model training is performed towards detection and classification along inclusion of feature extraction and feature selection mechanism [5]. Especially convolution layer of the model performs both feature extraction process through kernel function and feature classification process through softmax function along optimizing it layer through Adam optimizer in the fully connected layers. Further experimental and performance efficiency of model is computed using cross fold validation. Finally specified model provides 94.8% accuracy.

3. PROPOSED MODEL

In this section, architecture of the proposed deep learning approach entitled as Stacked Recurrent Neural Network with Long Short Term Memory Model towards detection and classification of the cyberbullying post to the user generated content collected from twitter dataset. Figure 1 represents the architecture of the proposed model along the other architecture elements towards preprocessing, experimental and performance analysis. Detailed analysis of each architectural elements of the model is as follows

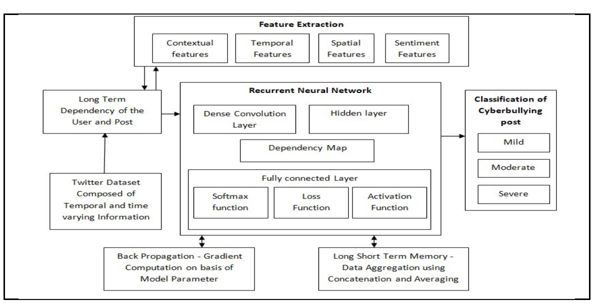

Figure 1: Architecture of RNN+LSTM towards Cyberbullying detection and Severity Classification

3.1 Dataset preprocessing

Dataset preprocessing phase composed of the multiple sub phase represented as noise removal, stop word removal, stemming, tokenization and token weighting. Preprocessing is performed on the raw data in form of user post to the user generated content in the twitter platform

3.1.1 Noise Removal

Noise Removal phase is sub phase of the preprocessing phase composed of URL removal, hash tag removal and punctuation removal mechanism. URL removal mechanism removes URL included in the post to the user generated content, hash tag removal mechanism removes Hash tag included along the post and finally punctuation removal mechanism removes punctuation marks such as comma, full stop, brackets, colon, hyphen etc

3.1.2 Stop Word Removal

Stop word removal phase is another sub phase of the preprocessing phase composed of the meaningless and frequently occurring words removing mechanism occurring in form of noun, pronouns articles, preposition and conjunctions. Especially stop words in the post to the user generated content have less semantic value. Thus, stop words like A, AND , THE , LIKE , BUT is removed from the post to the user generated content in the natural language processing.

3.1.3. Stemming or lemmatization Mechanism

Stemming mechanism is a sub phase of the of the data preprocessing phase composed of the text normalization mechanism. In Natural Language processing, stemming or lemmatization process reduces word length of its base word in the post. In particular base word which frequently affixed with suffixes and prefixes will be removed using lemmatization[6]. Lemmatization performed on basis of word context and morphological analysis of the base word.

3.1.4 Tokenization

Tokenization is a sub phase of the of the data preprocessing phase composed of the word splitting mechanism. In Natural Language processing, it is used to transform the unstructured text into structured text. Further tokenization can be enabled as word based tokenization, sentence based tokenization and character based tokenization. Tokenization partitions the text into significant tokens. Significant token can be identified as word, sub-words and characters. In this work, word based tokenization is incorporated which splits the post as word and those words were represented as individual tokens.

3.1.5 Token Weighting Scheme-TF-IDF

Token weighting scheme to tokenized word obtained from the cyberbullying post is performed using Term frequency and Inverse document frequency mechanism. It is a process of the assigning weights to the tokens on basis of its importance within a context. Especially it focuses on the significant tokens to the context and it helps mitigating the bias of the neural network model for various natural languages processing task. Further Term Frequency[7] is employed in this work as weighting scheme and it is defined as no of occurrence of the word token to the specified context of the particular text . Term frequency based weighting approaches to token is as follows

$$TF(T,P) = \frac{No\ of\ Occurence\ of\ the\ word\ T\ appear\ in\ the\ post\ P}{Total\ no\ of\ token\ in\ post} \quad (1)$$

Where T is word token and P is post

$$IDF(T,C) = \log\left[\frac{Total\ Number\ of\ post\ to\ the\ specified\ collection\ C}{Number\ of\ post\ containing\ word\ T}\right] \quad (2)$$

Where T is word token and C is data collection containing multiple post

$$TF\text{–}IDF = TF(T,P) * IDF(T,C) \quad (3)$$

Weighted Token is gathered as bag of words which contains weighted token of the post and entire dataset collection. Weighted Token is represented as fixed length sparse vector and it is projected to the proposed deep learning model.

3.2 Long Short Term Memory

Long Short Term Memory requires sequence of the embedding to capture temporal relationship as it process sequence of dense vectors instead of fixed length sparse vector. Thus it requires transformation of the sparse vector to dense vector to get processed in LSTM Model. In order to perform transformation, word2vector embedding has to be applied to sparse vector. Further word embedding is applied to TF-IDF computation again to obtain weighted word embedding [8]. Finally weighted word embedding is processed in internal cell states of the LSTM.

3.2.1 Gating Mechanism

LSTM model uses gating mechanism to control the sequence of the input through input gate, forget gate and output gate. Gating mechanism of the LSTM model focuses on most relevant information either from immediate past or from distant past. In particular, input gate process the input sequence and store the cell state information in the memory as hidden state. Further forget gate process the information from hidden states and identifies which information has to be eliminated. Finally output gates process the information from the hidden state to regulate the information on generating the current output. Gate computation is as follows

Input gate $I_G = \sigma(W_i[h_{t-1},x_t]+b_i$ (4)

Where W_i is the Weight matrix, σ is sigmoid function b_i is bias and h_{t-1} is the hidden state of the previous time stamp and x_t is weighted word embedding.

Forget gate $F_G = \sigma(W_f[h_{t-1},x_t]+b_f$ (5)

Where Wf is the Weight matrix, σ is sigmoid function bf is bias and ht-1 is the hidden state of the previous time stamp and xt is weighted word embedding.

Output gate $O_G = \sigma(W_o[h_{t-1},x_t]+b_o$ (6)

Where W_0 is the Weight matrix, σ is sigmoid function b_0 is bias and h_{t-1} is the hidden state of the previous time stamp and x_t is weighted word embedding.

Candidate Cell state C_t =tanh(($W_c[h_{t-1},x_t]+b_c$ (7)

Where W_c is the Weight matrix, σ is sigmoid function b_c is bias and h_{t-1} is the hidden state of the previous time stamp and x_t is weighted word embedding.

Short Term dependency

Short Term Dependency is a relationship between the sequences of the input in form weighted word embedding. Short term dependency is highly significant as it analyze the immediate context within the sequence. Further it is effective in predicting the dependency in the time series data. State update of the short term dependency is mentioned as follows

Cell State $C_t = F_t * C_{t-1} + i_t * \hat{c_t}$ (8)

Hidden State $h_t = O_t * \tanh(C_t)$ (9)

Long Term Dependency

Long Term Dependency is used to learn and retain information from earlier time steps in the sequence to the later outputs. Long term dependency is highly significant as it analyze the present input sequence on the context from distinct past inputs.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

State update of the long term dependency is mentioned as follows

Cell State $C_t = F_t * C_{t-1} + i_t * \hat{C}_t$ 10)

Hidden State $h_t = O_t * \tanh(C_t)$ (11)

3.3 Recurrent Neural Network

Recurrent Neural Network process the cell state information of the LSTM model in the multiple layer of the model to obtain the desired results. In this model, convolution layer process the cell state information of the LSTM model containing the weighted word embedding to extract the contextual, temporal, spatial and sentimental features[9]. Those features were processed in the hidden layer which is connected with LSTM model towards feature aggregation and it representation in form of dependency map.

Table 1: Hyperparameter tuning of the Stacked RNN+LSTM

Hyperparameter	Value
Learning Rate	10^{-6}
No of Epoch	60
Activation function of LSTM	Sigmoid
Activation function of RNN	ReLU
Softmax function	SVM
Loss function	Cross Entropy

Next dependency map is processed in the fully connected later of the model containing activation function to linearize the dependency map from output gate of the LSTM and softmax function uses those features to detect and classify the cyberbullying post into different severity classes. Table 1 represents the hyperparameter tuning of the proposed stacked recurrent neural network with long short term memory model[10].

3.3.1 BERT Model- Convolution Operation

Weighted word embedding is applied to Bidirectional Encoder Representation from Transformer model. Transformer model uses convolution layer to perform Bidirectional Encoder Representation operations. In this layer, weighted word mapping is mapped to high dimensional vector representation through pre-trained vocabulary and pre-trained embedding for behavior context. Further contextual embedding and behavior embedding of word is obtained as contextual relationships and behavior relationship through filter mechanisms. It generates contextual, temporal, spatial and sentimental features to the weighted word embedding.

F_f (contextual feature) = filter (weighted word embedding | Pre-Trained Vocabulary) (12)

F_f (Semantic feature) = filter (weighted word embedding | Pre-Trained behavior) (13)

3.3.2 Hidden Layer

Hidden layer of the model extract the contextual, temporal, spatial and sentimental features from the convolution layer of the transformer model and process it to generate the dependency map on association with LSTM model. LSTM model provides cell states and gating mechanism to process the features in the hidden layer. Finally hidden layer generates the dependency map to the features extracted from convolution layer[12].

Feature Aggregation F_A = [F_f(Contextual feature) F_f(Sematic feature)+ F_f(Spatial feature) +F_f(Temporal feature) (14)

Dependency Vector = Feature aggregation at specified time stamp (15)

3.3.3 Fully Connected layer

Fully connected layer of the model composed of activation function, softmax function and loss function. Operation of each function is as follows

Activation Function

Activation function of the model uses ReLU approach[13] to linearize the dependency vector from output gate of the LSTM. Especially linear function process the dependency vector to generate linear dependency vector. Activation function is represented as

A_f = ReLu(Dependency Vector (16)

Softmax function

Softmax function of the model uses support vector machine classifier function[14] to predict and classify the dependency vector into the severity classes. In this function, dependency vector is transformed into the fixed length vector. Further fixed length vector is transformed into hyperplane. Finally decision boundary is set to the hyperplane in terms of the class labels to predict and classify the features according to it . Classification function is represented as

$C_f = W_t * x_t + \text{bias}$ (17)

Where W_t is the weight vector, X_t is the word vector

Classification rule

$Y = \text{sign}(c_f)$ (18)

If $c_f < 0$ → Mild Cyberbullying Class

If c_f between 0 and 1 → Moderate Cyberbullying Class

If $c_f > 0$→ Severe Cyberbullying Class

Loss function

Loss function is used to regularize the model against the overfitting and underfitting issues. In this work, cross entropy is used employed as loss function. Cross entropy function[14] to the model is represented as

$$L_f = [y \log(p)+(1-y)\log(1-p) \qquad (19)$$

Cross entropy function eliminate the overfitting and underfitting issues through hyperparameter optimization with weight updates and early stopping criteria to each layer to increase the regularization strength[15]. Algorithm 1 mentions the process of the proposed architecture

Algorithm: Stacked RNN+LSTM
Input : Twitter dataset containing Post or opinion to user generated context Output : Cyberbullying classes{ Mild(Humilating), Moderate(harassing), Severe(threatening)} Process() Phases_ Pre_Process() N_R = Noise Removal_URL (User opinion) | Noise Removal_Hash tag (User opinion) | , Noise Removal _Punctuation(User opinion) S_w = Stop word Removal (N_R) S_T = Stemming (S_w) W_T = Word Tokenization (S_T) T_w = Token Weighting_TF-IDF(W_T) Long Short Term Memory () Weighted Word Embedding $W_w(E)$ = Sequential Embedding(Weighted Token) Initial Hidden State = Sequence Weighted Word Embedding vector Gating Mechanism () Input gate _Sequence Weighted Word Embedding vector Forget gate _ Sequence Weighted Word Embedding vector Output gate _ Sequence Weighted Word Embedding vector Candidate Cell state _ Sequence Weighted Word Embedding vector Recurrent Neural Network() Convolution layer F_E(contextual)=Filter (Weighted Word Embedding |Pre Trained Vocabulary) Convolution layer F_S(Sentimental)=Filter (Weighted Word Embedding |Pre Trained behavior) Hidden layer () Feature aggregation $F_A = F_F + F_S + F_T + F_L$ D_V = Dependency Vector at each time stamp Fully Connected layer() Activation function A_f = ReLu(D_v) Loss function C_E = Cross entropy (A_f) Softmax function C_f = Class (A_f) Cyberbullying Class = {Mild (Humiliating), Moderate (harassing), Severe (threatening)

4. EXPERIMENTAL ANALYSIS

In this section, experimental and performance analysis of the stacked recurrent neural network with long short term memory model is performed using the twitter dataset composed of the user opinion (post) to the user generated content in the python environment. Initially extracted dataset collection is partitioned into to train data and test data towards model training and model testing. Model training is performed using Scikit and Tensor flow framework[16] from python environment on its architectural elements of the proposed architecture through training data and model testing of the proposed architecture is performed using confusion matrix on cross fold validation of the testing data. Confusion matrix provides detailed analysis of the model against its classification performance. Each cell of the matrix represents the sample counts of the actual class and predicted class

4.1 Dataset Description: Twitter Dataset

Twitter dataset is social media data composed of the 47000 labeled tweets specifically classified according to various categories related to cyberbullying such as age, ethnicity, gender, religion and other factors of cyberbullying. Dataset is extracted between years March 2023 to August 2024. Further it includes various type of cyberbullying such as harassing, humiliating and threatening [17]. Furthermore dataset composed of user specific features on the opinions like emotion and sentiments. Finally distinct categories of the cyberbullying post content on the various contexts along its attached sentiments from the cyberbullying user are included.

4.2 Performance Analysis

Performance analysis of the model is computed using performance metric such as precision, recall and accuracy. Performance metric utilizes the values from the cell of the confusion matrix which determines the classification ability of the model. Cells of the confusion matrix contains True positive which mentions correctly predicted positive cases, True Negative mentions correctly predicted negative cases, False positive mentions incorrectly predicted positive cases and False negative

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mentions incorrectly predicted negative cases[18]. Figure 2 represents the confusion matrix of the model. Thus performance computation of the classification model is as follows

Figure 2: Confusion Matrix

4.2.1. Precision Analysis

Precision analysis is defined as ratio of the correctly predicted positive dependency features to total number of dependency features extracted as predicted positive dependency features on processing user opinions[19]. Precision analysis is as follows

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + Fals\ \ Positive} \quad (20)$$

Figure 3 illustrates the precision analysis of the proposed model against the multiple categories of the cyberbullying

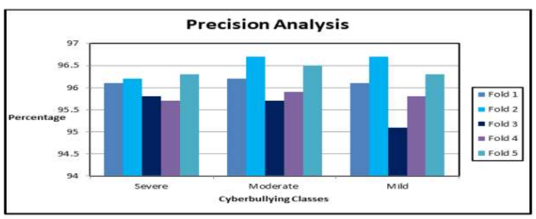

Figure 3: Precision Analysis

4.2.2. Recall Analysis

Recall analysis is defined as ratio of predicted positive dependency features to the cyberbullying classes among the total number of dependency features extracted positive features from the user opinions. Recall analysis is as follows

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + Fals\ \ Negative} \quad (21)$$

Figure 4 illustrates the precision analysis of the proposed model against the multiple categories of the cyberbullying

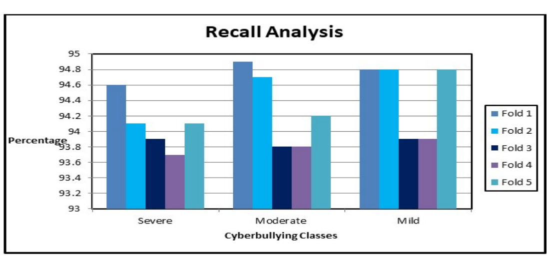

Figure 4: Recall Analysis

4.2.3 F Measure

F measure is defined as harmonic mean of the precision and recall as it is uses false positive and false negative cases on processing dependency features of the user opinion to the cyberbullying classes. Especially it performs well with imbalanced dataset. F measure is as follows

$$\text{F Measure} = \frac{2*(precision*Recall)}{Precision+Recal} \quad (22)$$

Figure 5 illustrates the precision analysis of the proposed model against the multiple categories of the cyberbullying

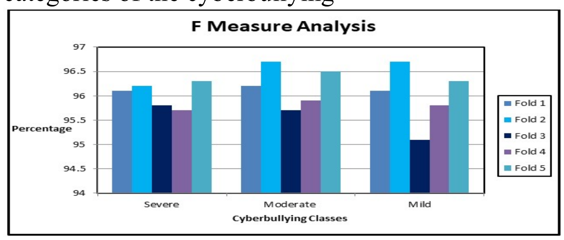

Figure 5: F- Measure Analysis

On analysis of the performance of the proposed methodology against the 5 fold validation of the model, it is proved that model attains average precision of 98.2% to moderate cyberbullying, 98.1% to mild cyberbullying and 98.4% for severe cyberbullying prediction categories, average recall of 96.1 % to moderate cyberbullying, 95.9% to mild cyberbullying and 96.4% for severe cyberbullying prediction categories and finally average f measure of 98.1 % to moderate cyberbullying, 97.9% to mild cyberbullying and 98.4% to severe cyberbullying prediction categories. Table 2 mentions performance analysis of the proposed classification model on cross fold validation.

Table 2: Performance analysis of the Stacked RNN+LSTM classification model on cross fold validation

Validation Data	Class	Precision	Recall	F Measure
Fold 1	Severe	96.1	94.8	96.2
	Moderate	96.2	94.7	96.1
	Mild	96.4	94.2	96.4
Fold 2	Severe	96.1	94.1	96.2
	Moderate	96.3	94.2	96.7
	Mild	96.5	94.8	96.2
Fold 3	Severe	95.1	93.9	95.9
	Moderate	95.6	93.7	96.1
	Mild	95.9	93.8	95.9
Fold 4	Severe	95.8	93.2	95.9
	Moderate	95.7	93.7	95.8
	Mild	95.8	93.9	96.1
Fold 5	Severe	96.1	94.1	96.2
	Moderate	96.2	94.6	96.8
	Mild	96.4	94.5	96.7

Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3. Performance Comparison

Performance comparison of the proposed stacked Recurrent Neural Network with Long Short Term Memory model is compared against conventional approaches like Recurrent Neural Network using Dolphin Echolocation Algorithm, Convolutional Neural Network models like ResNet and DenseNet and Transformer Based model on processing twitter dataset containing user opinions to the user generated texts. Performance comparison of the model has been carried out on basis of the performance accuracy using single fold test data on binary classification. Figure 6 mentions the performance comparison of the deep learning model towards opinion classification against cyberbullying and non cyberbullying.

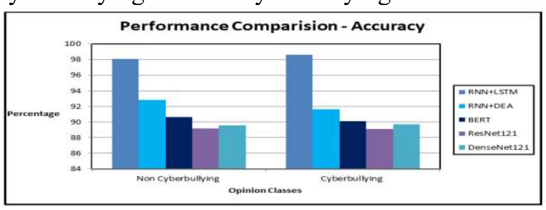

Figure 6: Performance Comparison of Deep learning model towards cyberbullying detection

On performance comparison of the deep learning model towards cyberbullying detection, it is proved that proposed stacked recurrent neural network with long short term memory model produces increased performance accuracy of 98.1% for non cyberbullying and 98.6% for cyberbullying Table 3 mentions performance comparison of the deep learning model towards cyberbullying detection and classification

Table 3: Performance Comparison of Deep learning model on cyberbullying classification

Classes	RNN+LSTM – Proposed	RNN +DEA	BERT	ResNet 121	DenseNet 121
Non Cyberbullying	98.1	92.8	90.6	89.2	89.6
Cyberbullying	98.6	91.6	90.1	89.1	89.7

Conventional model like Recurrent Neural Network[20] using Dolphin Echolocation Algorithm[21] produces performance accuracy of 92.8% for non cyberbullying and 91.6% for cyberbullying , Convolutional Neural Network models like ResNet[22] produces performance accuracy of 89.2% for non cyberbullying and 89.1% for cyberbullying and DenseNet[24] performance accuracy of 89.6% for non cyberbullying and 89.7% for cyberbullying and Transformer Based model like BERT[25] produces performance accuracy of 90.6% for non cyberbullying and 90.1% for cyberbullying. Finally it is proved that proposed model achieves improved detection and classification results towards identifying and classifying the various categories of the cyberbullying instances.

CONCLUSION

In this paper, a stacked recurrent neural network with long short term memory model is designed and implemented towards detecting and classifying the severity of the cyberbullying post as mild, moderate and severe. Initially, dataset is employed pre-processing phase with noise removal sub-phase mention URL removal, punctuation removal and hash-tag removal along sub phase like stop word removal , stemming , tokenization and token weighting. Weighted word token has projected to the LSTM model to transform the sparse word token vector into sequence word token vector as it supported to determine the short term and long term feature dependency using gating mechanism. Further recurrent neural network incorporates the LSTM model to store state information and it process the sequence word token vector in the convolution layer of the transformer model to obtain the contextual , sentiment, spatial and temporal features. Those features were processed in the hidden layer towards aggregation and it is stored in memory state of the LSTM to determine the dependencies. Dependencies vector is processed in the fully connected layer containing activation function to linearize the dependency feature , loss function reduce the overfitting and underfitting issues and finally softmax function to classify the dependency feature into cyberbullying severity classes as mild, moderate and severe. Finally experimental and performance analysis of the model proves that proposed model produces 98.1 percent accuracy against conventional approaches.

REFERENCES:

[1] S. Neelakandan, R. Annamalai, S. J. Rayen, and J. Arunajsmine, ‘‘Social media networks owing to disruptions for effective learning,’’ Proc. Comput. Sci., vol. 172, pp. 145–151, Sep. 2020.

[2] J. F. Hair and M. Sarstedt, ‘‘Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing,’’ J. Marketing Theory Pract., vol. 29, no. 1, pp. 65–77, Jan. 2021.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, A. Mohamed, and I. Ashraf, ''Water quality prediction using KNN imputer and multilayer perceptron,'' Water, vol. 14, no. 17, p. 2592, Aug. 2022

[4] A. John, A. C. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, and K. Hawton, ''Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review,'' *J. Med. Internet Res.*, vol. 20, no. 4, p. e129, Apr. 2018

[5] J. Wang, K. Fu, and C.-T. Lu, ''SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection,'' in Proc. IEEE Int. Conf. Big Data, Dec. 2020, pp. 1699–1708

[6] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, ''Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT,'' Information, vol. 14, no. 8, p. 467, Aug. 2023

[7] S. M. Fati, A. Muneer, A. Alwadain, and A. O. Balogun, ''Cyberbullying detection on Twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction,'' Mathematics, vol. 11, no. 16, p. 3567, Aug. 2023

[8] M. Karim, M. M. S. Missen, M. Umer, S. Sadiq, A. Mohamed, and I. Ashraf, ''Citation context analysis using combined feature embedding and deep convolutional neural network model,'' Appl. Sci., vol. 12, no. 6, p. 3203, Mar. 2022

[9] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, ''Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques,'' IEEE Access, vol. 6, pp. 59475–59485, 2018

[10] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, ''An application to detect cyberbullying using machine learning and deep learning techniques,'' Social Netw. Comput. Sci., vol. 3, no. 5, pp. 1–13, Jul. 2022

[11] A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan, and M. Prasad, ''Identification and classification of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting,'' in Proc. Int. Conf. Neural Inf. Process. Cham, Switzerland: Springer, 2020, pp. 113–120

[12] S. Paul, S. Saha, and J. P. Singh, ''COVID-19 and cyberbullying: Deep ensemble model to identify cyberbullying from codeswitched languages during the pandemic,'' Multimedia Tools Appl., vol. 82, pp. 8773–8789, Jan. 2022

[13] V. Vijayakumar, D. H. Prasad, and P. Adolf, ''Multi-input deep learning algorithm for cyberbullying detection,'' *Int. J. Res. Eng. Appl. Manag.*, vol. 7, no. 5, 2021

[14] B. A. H. Murshed, S. Mallappa, O. A. M. Ghaleb, and H. D. E. Al-ariki, ``Ef_cient Twitter data cleansing model for data analysis of the pandemic tweets," in Studies in Systems, Decision and Control, vol. 348. Springer, 2021

[15] H. M. Abdulwahab, S. Ajitha, and M. A. N. Saif, ``Feature selection techniques in the context of big data: Taxonomy and analysis," Appl. Intell., Jan. 2022

[16] D. Yu, Y. Yang, R. Zhang, and Y. Wu, Knowledge Embedding Based Graph Convolutional Network, vol. 1. New York, NY, USA: Association for Computing Machinery, 2021

[17] S. Min, Z. Gao, J. Peng, L. Wang, K. Qin, and B. Fang, ``STGSN A spatialtemporal graph neural network framework for timeevolving social networks," Knowl.-Based Syst., vol. 214, Feb. 202

[18] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, ''Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review,'' Expert Syst. Appl., vol. 167, Apr. 2021

[19] M. Fortunatus, P. Anthony, and S. Charters, ''Combining textual features to detect cyberbullying in social media posts,'' Proc. Comput. Sci., vol. 176, pp. 612–621, Jan. 2020

[20] M. N. Turliuc, C. Măirean, and M. Boca-Zamfir, ''The relation between cyberbullying and depressive symptoms in adolescence. the moderating role of emotion regulation strategies,'' Comput. Hum. Behav., vol. 109, Jan. 2020

[21] A. Mumuni and F. Mumuni, ''Data augmentation: A comprehensive survey of modern approaches,'' Array, vol. 16, Dec. 2022

[22] J. M. Ortiz-Marcos, M. Tomé-Fernández, and C. Fernández-Leyva, ''Cyberbullying analysis in intercultural educational environments using binary logistic regressions,'' Future Internet, vol. 13, no. 1, p. 15, Jan. 2021,

[23] N. M. G. Dwi Purnamasari, M. A. Fauzi, I. Indriati, and L. S. Dewi, ''Cyberbullying identification in Twitter using support vector machine and information gain based feature

selection,'' Indonesian J. Electr. Eng. Comput. Sci., vol. 18, no. 3, p. 1494, Jun. 2020

[24] D. Elreedy, A. F. Atiya, and F. Kamalov, ''A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,'' Mach. Learn., vol. 2023, pp. 1–21, Jan. 2023

[25] R. Essameldin, A. A. Ismail, and S. M. Darwish, ''An opinion mining approach to handle perspectivism and ambiguity: Moving toward neutrosophic logic,'' IEEE Access, vol. 10, pp. 63314–63328, 2022