**Cyber Bullying Detection using convolutional neural network**

**Sagar Rajendra Dalal1, Prasad Dilip Pophalkar1, and Nitish Das2**

*Student, Dept. of Computer Science and Engineering, MIT School of Engineering, Maharashtra, India***1**

*Assistant Professor, Dept. of Computer Science and Engineering, MIT School of Engineering, Maharashtra, India***2**

**Abstract:**

The use of social media is widespread and it is very common for people of varying ages to have an account in social media. For the same reason, misuse of social media profiles is very easy. Cyberbullying or bullying through social media websites have been an ongoing issue that rose with creation of social medias. Since it has adverse effects on people’s psychological behaviour, emotions and well-being, timely detection and prevention of social media bullying plays an important role in the world of internet. Machine learning and Deep learning are the common approaches adopted for cyberbully detection. In this research paper, we take a dataset containing tweets from the popular social media website Twitter to find texts that are possible cases of bullying. We create a hybrid bully detection system named Stacking Algorithm by combining three machine learning algorithms – K Nearest Neighbour (KNN), Support Vector Machine (SVM) and Random Forest (RF) to detect cyberbully texts accurately. The texts are checked and classified into three categories – Not Bullying, Racism and Sexism. We also include a section of Convolutional Neural Network (CNN) which identifies the bully texts accurately. A GUI is designed to display the accuracy percentage of the hybrid system as well as CNN. A comparison is made between the accuracy measure of the two systems and it is concluded that CNN provides a more accurate prediction than the Stacking Algorithm.

**Introduction:**

Cyberbullying refers to bullying or harassing people across digital media such as mobile phones, laptops and computers. Social media bullying refers to the bullying through social media websites like Facebook, Instagram, Twitter and YouTube. It involves sharing personal information about someone across the internet, posting abusive comments, sending hurtful and threatening messages etc. This issue is more common among teenagers. However, with the rapid increase in the use of social media by older people, social media bullying has become a common problem among people regardless of age. In a recent survey conducted, 47% young people have received nasty comments in their social media profiles and 62% people have been sent nasty private messages. Although there are options to give complaints regarding bullying, 91% of the people who reported complaints said that no actions were taken. Social media bullying victims undergo mental health issues hence it is necessary to develop a system that can detect bullying accurately and precisely. Several researches have been done in this field mostly using machine learning and deep learning techniques. Machine learning algorithms like Random Forest, Support Vector Machine, Naïve Bayes and KNN are some of the common algorithms used. Similarly, Convolutional Neural Network (CNN) is a deep learning approach adopted to detect cyberbullying. In previous works related to the topic, several attempts have been made to improve the existing algorithms through different methods. One such method is introducing new features to the algorithm used. Personal details of a user in social media, their popularity, number of followers and following etc. are few examples of features. These features have been mixed and matched in different ways to improve accuracy of the algorithm and to find the combination of features that produces best accuracy. Also, several studies were conducted to compare the performance of different ML algorithms and it could be concluded that SVM is the best technique out of all. Along with consideration of textual cyberbullying, visual cyberbullying was also studied. CNN is an algorithm suitable for finding bullying through images or videos. But very few attempts have been made to experiment in this field and there is still lots of scope to improve in this area. Other attempts mainly included bully detection in languages other than English such as Dutch, Bangla etc. and attempts to use the same algorithm in different languages.

In machine learning, there is a Training Dataset and Testing Dataset. In this research, dataset is taken from the social media website Twitter. The data which is in an unstructured form is cleaned by correcting spelling mistakes, removing stop words and forming tokens of words. 80% of that data is used to train the algorithm and 20% data is used to test the algorithm. The output of this research categorizes data into three, namely, non-bully, racism and sexism. When extracting data from social media it may contain a mixture of words and characters such as special characters, words in different language and alphanumeric characters. The data is mostly unstructured. This makes it difficult for traditional bully detection methods to detect bully texts precisely. Therefore, machine learning has emerged as the best approach to detect social media bullying as it is an automatic detection method. Hence, we propose a system which combines top three ML algorithms – SVM, RF and NB to form a hybrid algorithm called Stacking algorithm. Also, an analysis is done using CNN to check the accuracy in predicting bully texts. The data is first fed into CNN. It produces a Confusion matrix which shows the accuracy percentage of texts in each of the three categories. It also displays a text box where you can manually enter any data from the dataset to see the category to which it belongs to. After CNN is run, Stacking Algorithm starts to run and it follows the same steps to display the number of tweets falling under the categories.

Hence, the aim of this research is to:

1. Design a hybrid algorithm which combines the three best machine learning algorithms to analyse data from Twitter and provide accurate cyber bully prediction.
2. Compare the performance of the hybrid system with the existing CNN algorithm to check their accuracies in prediction.

**Literature Survey:**

Rui Zhao and Kezhi Mao [1] used a new representation learning method have been proposed to tackle this problem. This method is Semantic-Enhanced Marginalized Denoising AutoEncoder (smSDA) developed via semantic extension of the popular deep learning model stacked denoising autoencoder. the semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. This method is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text. This paper helps learn the denoising and autoencoding thus proved to be useful for the more efficient representation of the data.

ElahehRaisi and Bert Huang [2] proposed a weakly supervised machine learning method for simultaneously inferring user roles in harassment-based bullying and new vocabulary indicators of bullying. The learning algorithm considers social structure and infers which users tend to bully and which tend to be victimized. The weak supervision is in the form of expert provided small seed of bullying to address the elusive nature of cyberbullying using minimal effort and cost, and the algorithm uses a large, unlabeled corpus of social media interactions to extract bullying roles of users and additional vocabulary indicators of bullying. The model estimates whether each social interaction is bullying based on who participates and based on what language is used, and it tries to maximize the agreement between these estimates, i.e., participant-vocabulary consistency (PVC). Through this paper the role of PVC is studied and the information related to the detection of the bullying roles of user is learned.

P. Zhou, et. al. [3] proposed Attention-Based Bidirectional Long Short-Term Memory Networks (Att- BLSTM) to capture the most important semantic information in a sentence. The experimental results on the SemEval-2010 relation classification task show that this method outperforms most of the existing methods, with only word vectors. This paper proposes a novel neural network Att- BLSTM for relation classification. This model doesn’t utilize any features derived from lexical resources or NLP systems. The contribution of this paper is using BLSTM with attention mechanism, which can automatically focus on the words that have decisive effect on classification, to capture the most important se- mantic information in a sentence, without using extra knowledge and NLP systems. Through this paper the BLSTM with attention neural network and feature of BLSTM to classify the data more accurately is studied.

N. Srivastava, et. al. [4] explains the technique of dropout. The paper proves that dropout improves the performance of neural networks on supervised learning tasks in vision, speech recognition, document classification and computational biology, obtaining state-of-the-art results on many benchmark data sets. The key idea is to randomly drop units (along with their connections) from the neural network during training. During training, dropout samples from an exponential number of different “thinned” networks is taken. At test time, the effect of averaging the predictions of all these thinned networks can be thus approximated by simply using a single unthinned network that has smaller weights. This significantly reduces overfitting and gives major improvements over other regularization methods. This paper helps to understand the advantages and uses of dropout. The effects of dropout are can be studied.

A. Conneau, et. al. [5] Studied The fundamental idea of ConvNets is to consider feature extraction and classification as one jointly trained task. This paper presents a new architecture (VD-CNN) for text processing which operates directly at the character level and uses only small convolutions and pooling operations. The paper shows that the performance of this model increases with the depth sing up to 29 convolutional layers, and report improvements over the state-of-the-art on several public text classification tasks. ConvNets are namely adapted for computer vision because of the compositional structure of an image. Texts have similar properties characters combine to form n-grams, stems, words, phrase, sentences etc.This paper helps understand the use of deep convolution networks for the text classification and its advantages over RNN and LSTM.

S. Bhoir, et. al. [6] presented a comparative analysis of different word embedding models namely Continuous bag of words, Skip gram, Glove(Global Vectors for word representation) and Hellinger PCA (Principal Component Analysis). The models are compared on different parameters. The parameters are performance with respect to size of training data, basic over view, and relation of context and target words, memory consumption, supported classifier used and effect of changes in dimensionality. Word embedding turns text into numbers. This transformation has two important beneficial properties that is dimensionality reduction for efficient representation and contextual similarity for expressive representations. Thus, this paper proves useful to understand the benefits of various word embedding models and the comparison between them so as to select the best model.

E. Raisi, et. al. [7] presented the participant-vocabulary consistent model a weakly supervised approach for simultaneously learning the roles of social media users in the harassment form of cyberbullying and the tendency of language indicators to be used in such cyberbullying. The PVC can discover examples of apparent bullying as well as new bullying indicators in part because the learning process of PVC considers the structure of the communication network. It evaluates PVC on all social Medias platform data sets with both quantitative and qualitative analysis. The weekly supervised algorithm extrapolates from weak indicators to find possible instances of bullying in the data. Then the observable data from SMP’s is formalized. There is a model designed that checks for the entire conversation between the 2 persons using the built score and victim score. This model requires no extra space beyond the storage of vectors and raw data. Thus, this paper helps in studying the PVC technique.

H. Zeng, et. al. [8] used a visualization technique that has 4 linked view that’s helps to analyze learning parameters. These study uses AlexNet as the neural network architecture. It is necessary to get the insight of how the model parameters evolve from lower to higher accuracy so that we can improve the training process. The two main challenges in exploring the relationship between model parameters and performance i.e Scalability and Interpretability are solved here. The 4 view that the system provides are network architecture view, Difference distribution view, convolution operation view and performance comparison view. All the 4 view combined help to understand the insight of CNN clearly. Various parameters and activation values between two CNN snapshots are evaluated based on TFlearn framework. As training process of CNN leads to large number of parameters over time, this results in decreased performance. This paper helps to view the learning process of the CNN.

V. N. Kumar, et. al. [9] proposed that the effective representation of content is necessary for proper learning. This paper use naïve Bayes as the classifier for the content classification in email application it deals with the classification of spam words when massage is received and it is processed using feature set extraction method in which feature probabilities are found using NB and SVM are compared for precision factor .this paper just classifies the massage into cyberbullying .the denoised value for each word is calculated by grouping massage .this system alerts the system .it uses word embedding .technique which obtains bullying character automatically .the various modules use in this paper are GUI designing ,training dataset, classification and analyzing the twitter massages. A strong representation and learning of text massage are crucial for consistent detection system. The main method for the data extraction is web base mining technology.

Andrew M. Dal and Quoc V. Le [10] proposed supervised sequence learning model using CNN and LSTM. The semi-supervised learning is the combination of supervised and unsupervised learning where by using this the unlabeled data is proved to be more useful in improving the generalization of subsequent supervised model. The paper recommends the use of LSTM-RNN to be more useful than CNN and RNN for the purpose of data training using the proposed approach. The sequence autoencoder is used here to reconstruct the input sequence itself i.e. the original sequence. This paper uses LSTM due to certain benefits such as maintaining information ordering. This paper tests the semi-supervised method on five benchmarks to check the results using LSTM as the training method. This paper proves that CNN-LSTM is the better method than conventional CNN and gives better results than the previous methods for training unlabeled data.

K. Duan, et. al. [11] explains SoftMax combination for multicategory classification both oneversus-all and one-versus-one classifier. This paper explains how to efficiently extend binary classification method for multi-category classification. The paper also explains that most common approach to multi-category classification are binary-classifier based methods such as “one-versus-all” and “one-versus-one” that solves the multicategory classification problem. The one-versus-all method is usually implemented using a “winner-takes-all” strategy. Whereas the one-versus-all method is usually implemented using max wins voting strategy here the multicategory classification method is defined using these two classifiers through a SoftMax function. posteriori probabilities are obtained from the combination are used to do multicategory classification. This paper helps understand the advantages of multicategory classification and also methods to implement that using two methods in detail.

Q. Li, et. al. [12] proposed a new tweet sentiment classification approach using SSWE and WTFM produce classes based on the weighting scheme and text negation and a new text classification method. The method here is proved to be better than the SSWE and NRC techniques. In this the sentiment of tweet is polarized into 3 types. The paper suggests SSWE word embedding algorithm for data representation as it also do sentiment classification. The WTFM has 2 features i.e. negation feature and the tf.idf word weighing scheme. In the model here (SSWE + WTFM) the four features of WTFM concatenated with SSWE. The SSWE captures the semantic and syntactic feature and using original n-gram polarity of tweets it predicts 2-dimentional vector (f0, f1).

A. EI Adel, et. al. [13] Proposed Deep Convolutional Neural Network are used for the dropout and layer skipping. There is key advantage: rapid way to compute the feature using fast beta wavelet transform.the purpose intelligent dropout method.is based on a unit is efficiency and not randomly selected.it is possible to classify the image using efficient unit of earlier layer and skip the all hidden layer from the output layer. This paper proves that the FWT are the best or it is extracting the feature of input image.

S. Zhai, et. al. [14] Studied and proposed the context of search based online advertising. we used the recurrent neural network to map both queries and ads to real valued vectors with which the relevance of a given pair can be easily computed. we propose a novel attention score to different word location according to their intent importance .the vector output of a sequence in this computed by a weighted sum of the hidden state of the RNN at each word according their attention score. This paper proves that the RNN which allow us to model word sequence which is show to be of great importance to accurately capture the meaning of sequence.

I. Raid [15] explained development and advancement of technology in addition to bring a positive impact also introduced new problem when used inappropriately. This is often referred to as cybercrime. On of cybercrime is being lively at the moment is cyberbullying. social media is one of the for the development of cyberbullying .the research was carried out using data mining technique here there are several stages as data collection ,preprocessing, TF-IDF, weighting, data validation and classification using naive Bayes classifier. This paper prove that TF-IDF weighted and validation data using for the cross validation and then do classification.

K. Sahay, et. al. [16] explains that online bullying and aggression against social media user have grown abruptly. It affects more than half of young social media user recurrent implementation of insult detection using machine learning and natural language processing have very low recall rate. The research experimenting with different work process makes a robust methodology for extracting text, user experimenting with different methods the work process a robust methodology for extracting text, user work in certain ways to identify and classify bullying in the text by analyzing and network-based attributes studying the properties of bullies and aggressor and what feature distinguish them for regular user the NLP and machine learning are studied and evaluated for the task of identifying bullying comment in the dataset. This paper shows the training in machine learning model using supervised learning.

Several studies and researches have been conducted on finding out ways to detect cyberbullying, prevent them and take actions against them. Out of them, machine learning has proved to be the most effective technique in cyber bullying detection. Many improvements and innovations are brought to this field and in this section, we cover some of the major researches done in the field. One of the oldest studies trace back to the study done by Dadvar and Jong in [1] who suggested involvement of user information and their behaviour after harassment along with state of the art techniques which included only content of the conversation in social media sites for detecting cyberbullying. They suggested using personal information of users such as age and gender. They also suggested monitoring the posts of same users who got harassed in different social media websites. The above-mentioned features were added in bully detection process and it produced more accurate results. In [2], the authors have introduced the use of Fuzzy logic for detecting social media bullying. A genetic algorithm is applied on the input data which is basically the conversations taken from social media. This helps in early detection of cyberbullying before someone becomes the victim and it has proved to provide improved efficiency in cyber bullying detection.

There are also several papers that compare different approaches which detect cyber bullying to find out which method produces better accuracy. [3] discusses some common data pre-processing techniques such as tokenization, stop words removal, stemming, lemmatization, case folding and replacement of special symbols and compares algorithms used for bully detection such as SVM, Naive Bayes and J48. It was found that Support Vector Machine (SVM) gives the best and most accurate results. Social features were also taken into consideration in the algorithm for detection of bullying. Kelly and colleagues collected data from Formspring.me to identify bullying content through language- based bully detection method [4]. They used an Amazon web service to label the input data. The percentage of insults and curse words in the data were recorded and checked with machine learning algorithms J48, JRIP, IBM and SMO. 78.5% of cyber bullying posts were correctly identified using these algorithms and it was concluded that percentage of bad words produced more accurate results than count of bad words.

A different and interesting approach was initiated by Hao Li and team in 2016 [5]. It involved examining images posted on Instagram and the captions and comments associated with them to detect chances of cyberbullying. Several classifiers and feature sets were considered for the experiment which combined image content, their comments and contextual features. Image features were extracted using a convolutional neural network (CNN). It involved applying the results of a convolutional neural network which was already trained to the image pixels. Bullying in comments were classified using Bag of Words model. It was concluded that captions of pictures proved to be a significant predictor of cyber bullying. Similarly, in an experiment done by Dadvar and Eckert [6], they adopted deep learning techniques for cyber bully detection. The same models used previously in Formspring, Twitter and Wikipedia datasets were applied on data from Youtube to compare the results with output from conventional machine learning techniques. The deep learning models used were namely, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BLSTM) and BLSTM with attention. The results obtained concluded that Deep Neural Network (DNN) models outperformed the ML techniques and produced better and accurate results.

Another attempt made using deep learning was in 2018. It used a novel algorithm CNN-CB which eliminated the major steps of feature determination, feature extraction and selection. CNN-CB is based on convolutional neural network which uses the concept of word embedding, where similar words have similar embedding [7]. Hence, all bully words would have similar representations. It outperforms the traditional detection algorithms not only by eliminating major classification steps but also by producing an accuracy of 95%. An approach to automatically detect cyberbullying in social media including different types of bullying was done by students of universities in Belgium [8]. Signals of cyberbullying were detected in English and Dutch languages. The method involved the use of a linear SVM classifier which is used to detect signals related to cyberbullying. Another aim of this study was to find out whether this approach could be used for any languages although the experiment was done in English and Dutch. Precision, recall and F1(accuracy) scores were calculated. The maximum F1 score obtained for the combination of best three features were 64.26% and 61.20%, for English and Dutch, respectively. Similarly, for Recall, 57.19% and 66.40% and for Precision, 73.32% and 56.76% were obtained. Another approach related to linguistic difference was taken to detect cyberbullying texts in Bangla language [9]. Various machine learning algorithms like SVM, NB, J48 and KNN were checked to identify the best out of them. The experiment was done to classify the data into two classes- bullying or non-bullying. It was done in two phases. First phase consisted of checking only texts such as posts and comments in social media. The second phase consisted of considering user information along with textual features. In both phases, SVM showed the best results. It showed an accuracy of 95.40% in first phase. Adding user information also resulted in better performance of the algorithm.

In 2019, Vimala and colleagues researched on including psychological features of a person to machine learning algorithms to detect cyberbullying [10]. Tweets from Twitter were taken and three algorithms, namely, Naïve Bayes, Random Forest and J48 were used to classify the tweets into 4 categories. The psychological factors considered included personality, sentiments and emotions of a person. Naïve Bayes was the poorest performer among the three algorithms. Bringing psychological features were shown to improve cyberbully detection in J48 and Random Forest. J48 performed slightly better than RF and it produced an accuracy of 91.88%. It was also noted that factors like popularity of a person, number of followers and following also had an impact on detecting cyberbullying. In [11], an experiment was conducted to check data from Twitter. Log data was collected, preprocessed and weighted using TF IDF. The classification of data was done by NB classifier. Four algorithms were checked to see the accuracy. Logistic Regression showed the highest accuracy of 95.1%, followed by SVM with an accuracy of 94.6%, followed by KNN algorithm with 93.5% and Random Forest algorithm with an accuracy of 91.8%. A research done by Angelis and colleague [12] goes through previous studies done on ML algorithms that detect and prevent cyberbullying. Most of the researches focus on content-based features which means texts from social media posts and comments are used to predict bullying. Very few studies have focused on bullying based on visuals such as through images and videos. Similarly, very few studies took upon features like profile information from users, their followers etc. Three algorithms were discussed in this paper – Naïve Bayes, SVM and a deep learning algorithm CNN. Four metrices were calculated which included accuracy, precision, recall and F1 score. SVM is more commonly used and accurate than NB and it is used for identifying textual bullying while CNN is mainly used for identifying visual bullying.

A study done by Malpe and colleagues [13] discusses different researches performed in the field of detecting cyberbullying. It also described possible research gaps in the same field.

(Dalvi, et al., 2020)developed a system to detect cyber bullying using supervised form of machine learning algorithms. The system was based on identifying tweets from social media sites whether a particular tweet consist of any kind of hate speech. The algorithm used for detecting bullying texts were SVM and Naïve bayes algorithms. The accuracy obtained for SVM algorithm was about 71.53% while the accuracy obtained for Naïve Bayes was about 52.70%. The results obtained showed that the Support Vector Machine algorithm outperforms the Naïve bayes algorithm for a similar kind of work and for same dataset. Tweepy a twitter based API was used for extracting tweets from twitter.

John Hani et al. developed (Mounir, et al., 2019)a machine learning based architecture to detect cyber bullying from twitter. NLP based pre-processing was performed on the dataset:

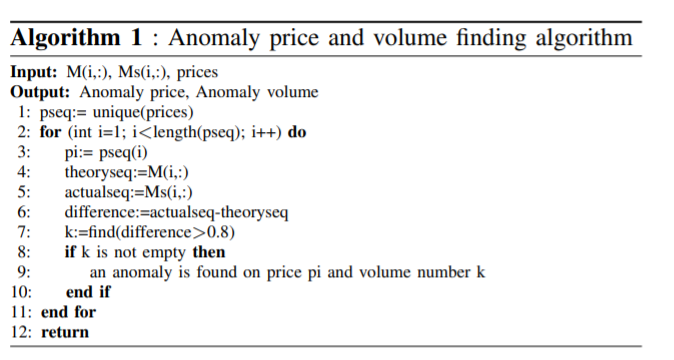
* Tokenization
* Lowering text
* Stop words and encoding cleaning
* Word correction

Feature extraction: Sentiment analysis and TFID algorithm was applied for extracting features for machine learning model.

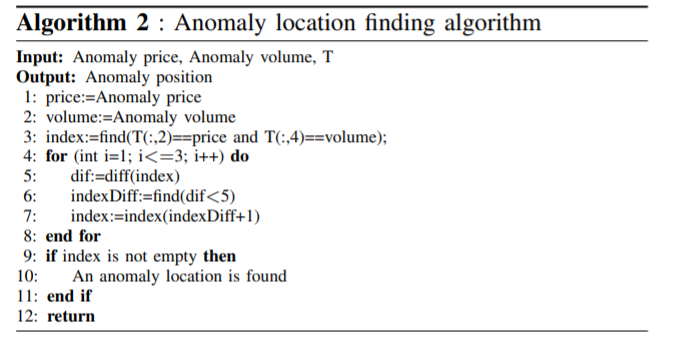
SVM and NN classifiers were used for making classifications. The accuracy obtained was about 92.8% and 90.33%.

In 2015 a thesis was developed by (Lei & Wang, 2015) which conducted a research on twitter data in order to predict stock market prices. NLP was performed on twitter data. Initially anomalies were used for making predictions for upward trend in in stock market. Tick by tick data was used on stock pricesinstead of the time series. The experiments for predicting stock market was initially conducted on some 200 tweets. The experiments done showed promising results. The result was successful in predicting upward trend in stock market.

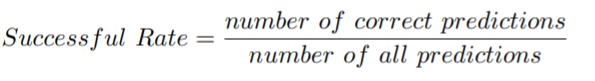
The algorithm proposed by (Lei & Wang, 2015) is:



A = {s, p, v} describes an anomaly record where stock index = s, anomaly price = p, anomaly volume = v. Another algorithm is used after all the anomalies get defined inorder to locate those anomalies on time during trade.



SVM classifier was used as machine learning model for making prediction. Plotting of graph was done based on the return of stock on an average basis. Same stocks were also measure using SVM algorithm for comparison purpose.



(Yin, et al., 2017)conducted a research on CNN and RNN algorithm. The thesis aim was to conduct a comparative study of CNN and RNN machine learning algorithms. In order to perform the comparison 7 different studies were done:

**Sentiment Classifier**: SST (Socher, et al., 2013) dataset was used for conducting test. Movie sentiment analysis was performed using both CNN and RNN algorithms for performing the study. Dataset was divided into training and testing set and final results were obtained after running the model. Accuracy measure was obtained for both the algorithm.

**Related Classification**: Related classification was performed on the dataset obtained by (Hendrickx et al., 2009). The result measure value used for comparison was F1-score. Sentences from dataset were labelled manually. The dataset was splitted for training and testing phase. Training dataset consisted of 8000 samples while testing part has 2717 samples.

**Textual Entailment:**The dataset used for conducting textual entailment based text was obtained from (Bowman, et al., 2015). The dataset consist of sentence having premise hypothesis pair labelled with a relation (entailment, contradiction, neutral). The accuracy was used as the measurement value. CNN and KNN algorithms were applied.

**Answer Selection**: The sentences present were open domain question answer in the dataset used for answer selection based test. Accuracy comparison was done for both CNN and RNN algorithms which were Question Relation Match (QRM), Path Query Answering (PQA) and Part-of-Speech Tagging. By performing these tests (Yin, et al., 2017) were able to deeply understand the difference in the working of RNN and CNN with respect to NLP.

Traditional studies of cyberbullying were largely on a macroscopic view. These were conducted by social psychologists and scientists. However, these studies were largely focusing on the statistics of cyberbullying and concentrated more on the psychological way to prevent it. One introductory work has been presented in which several NLP models such as BOW, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are applied to detect bullying signals in social media. Their result has verified the possibility of automatic cyberbullying detection.

A recent work of researchers in Massachusetts Institute of Technology proposed a novel methodology which is used to identify the bullying that happens over the domain of YouTube, i.e., the bullying which takes place in the form of YouTube comments are identified. Moreover, the proposed system is used to classify the YouTube comment into any of the following categories: sexuality, culture, intelligence and physical attribute. However, the outcome of the model was proven to be very less and provided ambiguity at times.

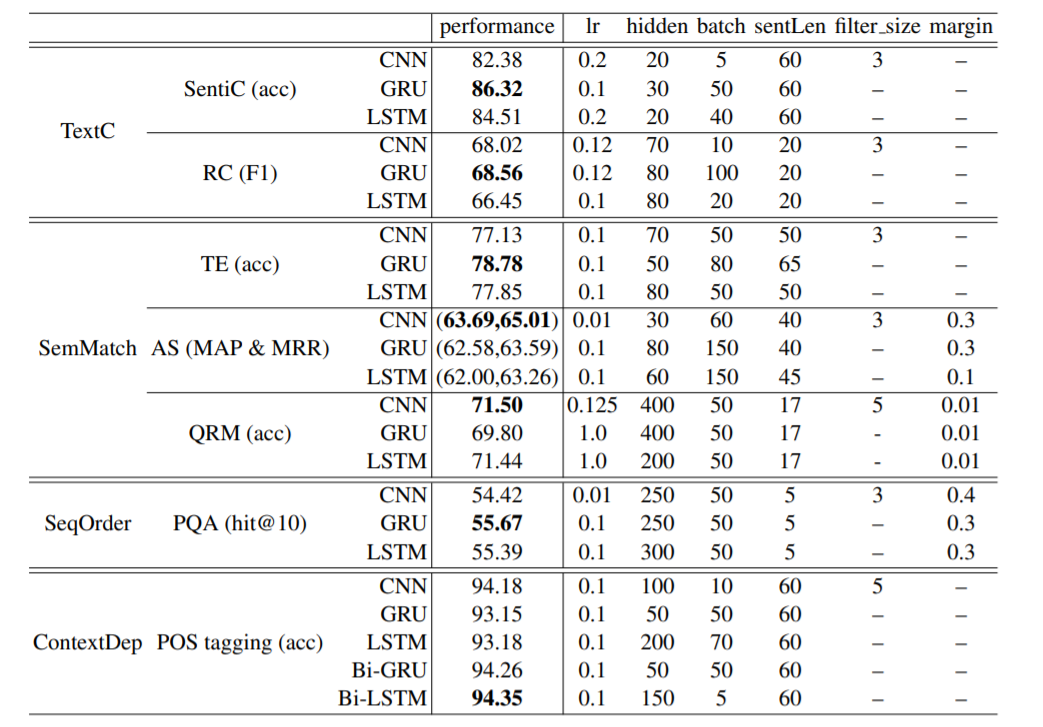
In other research work by Norton Online 2010 Malaysia for Detection of cyberbullying states that Malaysian children spend an average of 19 hrs a week on internet while that same also found that more than 80% of the children have been affected by negative comments and this is mainly done by their close friends or spouses. This work mainly focuses on the statistics of the cyberbullying.

Kontostathis et al. analyzed cyberbullying corpuses using the bag-of-words model to find the most common used terms by cyberbullies and used them to create queries capable of reaching a precision of 91.25% on average. Lempa et al. developed an Android application, embedded with two methods, to implement the cyberbullying detection. One method is built on a brute force search algorithm search for sensitive words and phrases within the text. The other method extracts words and phrases as seed words and detects cyberbullying online with keyword categorization and relevance matching. The top precision of both methods reaches 89% and 91%, respectively.

Regarding classifier design, researchers have tested various classifiers, including Naïve Bayes, C4.5 decision tree, Random forests, and SVM with different kernels on corpuses collected from popular social networks, such as Twitter and YouTube .Reynolds et al. found that both the C4.5 decision tree and 3-nearest neighbor classifiers can reach a recall of 78.5% on the text-based dataset collected from Formspring.me, a question and answer based social network. Bullying posts (positives) were duplicated 10 times to compensate for the imbalance within the data. However, this oversampling method is unreliable since it exaggerates the occurrence rate of the positive samples . On the Twitter dataset, Xu et al. showed that SVM with a linear kernel using unigrams and bigrams as features can achieve a recall of 79% and a precision of 76%. Other works are focused on ensemble methods such as cooperative and hybrid classifiers. Dadvar et al. introduced two approaches to combine machine learning methods and expert systems. The different combinations depend on which classifier’s output is used as the input of the other. An accuracy metric called the area under the curve (AUC) was used to evaluate their approach. The hybrid system made up of expert system and Naïve Bayes classifier, achieving their highest AUC score of 0.76. Mangaonkar et al. evaluated 15 cooperative classifier combinations, including heterogeneous, homogeneous, and selective cooperation with different parallelisms. These ensemble classifiers are extremely complex and tuning the hyperparameters is difficult.

For feature selection, various textual content based features, such as the basic bag-of-words and advanced sentiment prediction, were used as the input to classifiers Kasture took advantage of a psychometric feature analysis tool called Linguistic Inquiry and Word Count (LIWC) used it as a feature extractor. These features were used to train a variety of classifiers and the best performance reached 96.3% recall and 98.4% precision on Random Forests using 10-fold cross validation on the Twitter dataset . To detect the cyberbullying and cyberstalking in emails and messages, Ghasem et al. selected the 500 most informative words as the feature vector and achieved an F1 score of approximately 95% on SVM and a neural network classifier. Nahar et al. introduced a weighted TFIDF feature extractor and used LIBSVM with a linear kernel to detect cyberbullying content in three social networks: Kongregate, Slashdot, and Myspace. The experiment results show that their feature design significantly improved the performance of the baseline LIBSVM. For example, the recall jumped from 25% to 98% on the Myspace dataset. However, oversampling was used to handle the imbalance problem, which is not a useful method in real-world implementations .

In a research work Detection of cyberbullying is done by using naïve bayes technique. This method works fine for less data but for large data the outcome decreases. And also, this method gives best result in binary classification. One of the major problem with the existing methodology is that all the activities are focused more on the aftermath of the cyberbullying incident rather than an effective system to prevent the cyberbullying activities. Prevention is better than cure is the crux of our proposed architecture. Our paper targets on detecting the cyberbullying activities and classifying them into Cyberbully and non-Cyberbully which prevents the victims from facing the issues of cyberbullying and helps in taking preventive actions such as law enforcement, blocking or taking legal actions accordingly.



Performance comparison report for RNN and CNN algorithm

The above table displays the results obtained by the implementation of all the tests discussed above. In case of SentiC GRU KNN algorithm has the best performance. In case of experiments of QRM and AS CNN has the better accuracy. **In case of SeqOrder and ContextDep CNN has the worst performance. The results for ContextDep and SeqOrder are as expected: RNNs are very much suited for encoding order information (for PQA) and long-range context dependency (for POS tagging). The overall conclusion made was that the RNN outperforms CNN in wide range of task except when it comes to keyphrase detection and question-answer based sentences.**

**Related Work:**

According to the research conducted by (Jang, et al., 2019) it is clear that CNN can follow two kind of approach OnehotEncoder and Wordtovec for generation of vectors. The results obtained by the experiments conducted by (Jang, et al., 2019) showed that wordtovec boosts the overall accuracy of the system when compared with onehotencoder. Hence, wordtovec is preferred for this thesis.

According to the comparative study performed by (Yin, et al., 2017) between CNN and KNN algorithms, the results showed that RNN outperforms CNN when it comes to wide range of task, but CNN results are better with keypharases detection and question-answer based sentences which is exactly the kind of work that will get performed in this thesis. Hence, the thesis will use CNN algorithm for performing sentiment analysis.

Cyberbullying detection has a rapidly growing literature, even though researches addressing bullying are traced back to early 2010. The rich literature in this field can be divided into three categories: content-based, user-based and network based detection.

1. **Content-Based Detection** Among the first to tackle bullying in social media is [11], where a framework was built to incorporate Twitter streaming API for collecting tweets and then classifying them according to the content. Their work combined the essence of sentiment analysis and bullying detection. As a first phase, tweets are classified as being positive or negative and then they are further classified as positive containing bullying content, positive without bullying content, negative containing bullying content, and negative without bullying content. For the sake of classification, Naïve Bayes was implemented and resulted in a relatively high accuracy (70%). Another later research found in [12], incorporated statistical measures namely (TFIDF) and (LDA) along with topic models in order to extract relevance in documents. However, they did not rely on statistical measures only but extracted content features like: bad words and pronouns. Other researchers in [13], continued to pursue cyberbullying detection from content-based perspective however, they introduced new features like: emotions icon and dictionary of hieroglyphs. Their approach was tested using many learning algorithms: Naïve Bayes, SVM and J 48. And the best result was recorded with SVM achieving an accuracy of 81%. Another research [14], presented a prototype system to be used by organization members to monitor social network sites and detect bullying incidents. The approach followed relied on recording bullying words and storing them in a database and then incorporate Twitter API to capture tweets and compare their content to the bullying material recorded earlier. Beside the promising innovative idea in their work, this prototype system has not been implemented yet.
2. User-Based Detection Many researchers believed that user information like age and number of tweets could indicate potentiality to harm others. In [15], researchers incorporated user information like number of tweets, number of followers and number of followings into the detection process. Their total features -user based and others- resulted in good predictions with an accuracy of 85%. Similarly, in [14] they added user age as feature along with a history of a user as a feature. They assume that if a user bullied in the past it is more likely for him to engage in bullying again. They investigated the effect of adding user features and concluded that it advances the recall with 5%. User-based features were also adopted in [16], where they added user gender and age to the feature set. The assumption was that different gender use different language and the people from different ages have different writing styles. Moreover, a new user feature was incorporated which was the user location

**C.** Network based Detection An interesting perspective to cyberbullying detection studies the social structure of users. This starts by drawing network structure and deriving features from the graph. In [17], they focused on deriving features from social network graph. Features included: number of nodes indicating how large is the community and number of edges indicating how well connected is the community. Another research that addressed network based features is found in [12]. They used (Gephi) a graphical interface to visualise a user‟s connectivity based on the bullying posts. Then, they investigated the participants‟ role in the bullying, whether they are victims or predators

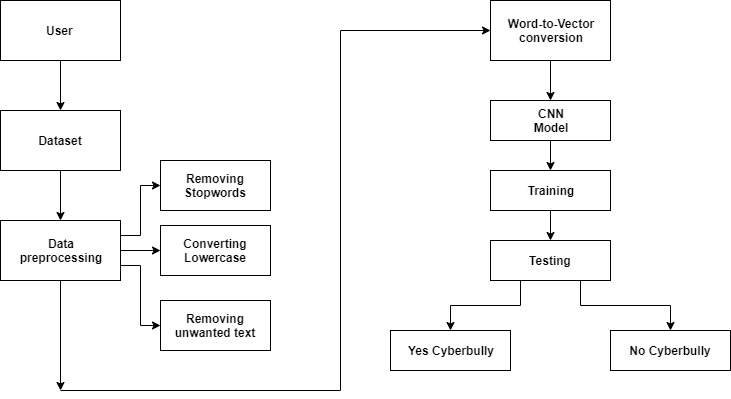
**Proposed System (Block Diagram):**

In our proposed system, the notion of CNN implementation is included. CNN is used with multiple layers which provide a process of iterative analysis over different layers to provide an efficient and accurate analysis. Inspired by the studies about the central nervous system of the mammals. A class of neural networks consist of significant number of layers of neurons, which are capable of learning by themselves is termed as deep learning. Deep learning in general consist of 3 layers,

• Input Layer

• Hidden Layer

• Output Layer



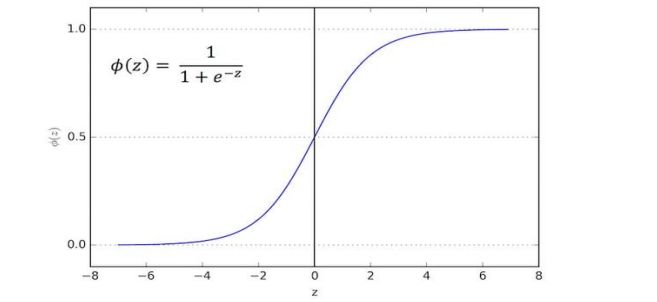
Our proposed model contains of a series of processes namely,

Data Pre-processing

Data pre-processing is cleaning of the data. It is the first and foremost step required in any process. It is the conversion of raw form of data into a required form of data for training the model in a proper manner. For example, in the raw form the data is “You look so ugly and fat #changethestyle”, after pre-processing the data is like “look ugly fat changestyle”. The pre-processed data takes out all the unwanted words like as, what, who, with, is, the etc. and special characters like @ () [] ?/; etc which are not required for training in the model. The data is separated into sentences and each sentence is made to make equal number of words by padding a common word which helps in uniformity of the data. Since the model accepts the data in the form of vector, the process makes the data into its lowercase format and converts that data into its vector form.

CNN Model Layers

The crux of the entire process depends upon the CNN layers used for processing. The main layers used in the model include Sequential Layer. The initial building block of keras is a model and the simplest model is called sequential model which consists of stack of neural network layers. The network is dense which means every node from each layer is connected with nodes from other layers. The perceptron is a single algorithm which takes the input vector x of m values as input and outputs either 1(yes) or 0(no) mathematically it is defined as f(x)=1 if wx+b>0 and f(x)=0 otherwise Perceptron is easy dealing with the small amount of data but in case of large data the perceptron is not helpful. That is, it cannot help in learning data. Since perceptron gives value either 0 or 1 the graph produced by it is discontinuous. We need something different and smoother. We need a function that progressively changes from 0 to 1 without any discontinuity.



Activation function can be of many types like sigmoid, ReLu etc. Sigmoid function is defined as 1/ (1+e− x) and can be used to produce continues values. A neuron can use the sigmoid for computing the nonlinear function z=wx+b where w is the weight of the neuron and b is the biased value. Activation function ReLu known as Rectified linear unit is also one such activation function which gives smooth values with nonlinear functions. A ReLu is simply defined as f(x)=max (0, x). The function is zero for negative values and grows gradually for positive values. In our network we have converted the input text to a sequence of word indices. For that we have NLTK (Natural Language Toolkit) to parse the text into sentences and sentences to words. We could have used regular expression but statistical models for nltk are more powerful than regular expressions. After creating the sequential model, the word indices are fed into array of embedding layers of a set size (in our case the longest word sentence) The output of the embedding layer is connected to the 1D Convolutional layer. This is then pooled into a single pooled word by a global max pooling layer. This vector is then input to a dense layer which outputs a vector (2) (Yes cyberbully and No Cyberbully). A SoftMax activation will return a pair of probabilies. The following shows our network model:

Embedding->Convolution1D->GlobalMaxpooling1D->Dense

Model Prediction

Once we define our model, we must compile it so that it can be executed by keras backend. (either Theano or Tensorflow). The model. compile consist of OPTIMIZERS, LOSS FUNCTION, METRICS. Optimizers are used to update weights while we train our model.

Once our model is compiled it can be trained with the fit() function. The parameters used are:

• epochs- This is number of time model is exposed to training set.

• batchsize- This is number of training instances before optimizer performs a weight update.

• validationdata- This is the data that needs to be tested.

**Conclusion:**

A comparative study of algorithm is performed between hybrid ensemble and neural network algorithm in order to find out which style of machine learning approach is best suited to tackle the problem of cyber bullying on social media. We have seen that CNN based neural network works or performs better as compared to other neural network. Sub-classifiers of stacking algorithm are formed by different supervised machine learning algorithm that has performed better as per the literature review. From the analysis of results and comparative study performed for both algorithm mentioned in result section we can conclude that stacking based machine learning approach are better when compared with neural network based approach in order to detect cyber bullying on social media platforms. The accuracy obtained for stacking is approx. 83% and that for Convolution Neural Network is about approx. 80%.

**References:**

1. “Text classification using convolution neural networks. (2017).

2. Keras tutorial deep-learning in python.

3. B. Sri Nandhinia, J. (2015). “Online social network bullying detection using intelligence techniques.

4. K. Dinakar, R. R. and Lieberman, H. (2011). “Modelling the detection of textual cyberbullying.

5. K. Reynolds, A. K. and Edwards, L. (2011). “Using machine learning to detect cyberbullying.

6. Mohammed Ali Al-garadi\*, Kasturi Dewi Varathan, S. D. R. (2016).

7. “Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network.

8. Rui Zhao, Anna Zhao, K. M. “Automatic detection of cyberbullying on social networks based on bullying features.

9. V. Nahar, X. L. and Pang, C. (2013). “An effective approach for cyberbullying detection.

10. Whittaker, E.,. K. R. M. (2015). “Cyberbullying via social media.

11. Archer (2018) B. Sri Nandhinia (2015) Mohammed Ali Al garadi\* (2016)

12. Rui Zhao (Rui Zhao) K. Dinakar and Lieberman (2011) K. Reynolds and Edwards (2011) Whittaker (2015) V. Nahar and Pang (2013) lin (2017)

13. Tweet classification of sentimental analysis using keras in python

14. Deep Learning for detecting cyberbullying across social media platforms by S Agarwal A Awekar.

15. Detecting state of aggression in sentence By R potapova

16. Hate speech detection on Facebook (Blog)

17. Analytics Vidya (Website for python and CNN)

18. S. Salawu, Y. He, and J. Lumsden, “Approaches to Automated Detection of Cyberbullying : A Survey,” vol. 3045, no. c, pp. 1–20, 2017.

19. T. Wu, S. Liu, J. Zhang, and Y. Xiang, “Twitter spam detection based on deep learning,” Proc. Australas. Comput. Sci. Week Multiconference - ACSW ‟17, pp. 1–8, 2017.

20. MIT Technology Review, “These are the 10 breakthrough technologies you need to know about right now,” 2018. [Online]. Available: https://www.technologyreview.com/lists/technologies/2017/. [Accessed: 02-Mar-2018].

[21] K. Reynolds, A. Kontostathis, and L. Edwards, “Using Machine Learning to Detect Cyberbullying.”

[22] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, “Detection of Cyberbullying Incidents on the Instagram Social Network,” 2014.

[23] B. S. Nandhini and J. I. Sheeba, “Online Social Network Bullying Detection Using Intelligence Techniques,” vol. 45, pp. 485–492, 2015.

[24] K. Dinakar, B. Jones, C. Havasi, and H. Lieberman, “Common Sense Reasoning for Detection , Prevention , and Mitigation of Cyberbullying,” vol. 2, no. 3, 2012.

[25] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[26] F. Chollet, “Keras,” GitHub, 2015. [Online]. Available: https://github.com/fchollet/keras. [Accessed: 26-Aug-2018].

[27] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” J. Mach. Learn. Res., vol. 3, no. Feb, pp. 1137–1155, 2003.

[28] N. Majumder, A. Gelbukh, I. P. Nacional, and E. Cambria, “Deep learning-based document modeling for personality detection from text,” IEEE Intell. Syst., vol. 32, no. 2, pp. 74–79, 2017.

[29] A. Géron, Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.

empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[26] F. Chollet, “Keras,” GitHub, 2015. [Online]. Available: https://github.com/fchollet/keras. [Accessed: 26-Aug-2018].

[27] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” J. Mach. Learn. Res., vol. 3, no. Feb, pp. 1137–1155, 2003.

[28] N. Majumder, A. Gelbukh, I. P. Nacional, and E. Cambria, “Deep learning-based document modeling for personality detection from text,” IEEE Intell. Syst., vol. 32, no. 2, pp. 74–79, 2017.

[29] A. Géron, Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.