2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART) | 978-1-6654-8734-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMART55829.2022.10047054

# Deep KNN Based Text Classification for Cyberbullying Tweet Detection

M. Nisha[1] and Dr. J. Jebathangam[2]
[1]*Research Scholar, Department of Computer Science, VISTAS, Chennai*
[2]*Associate Professor, Department of Information Technology, VISTAS, Chennai*
*Email: [1]manikantnisha23@gmail.com, [2]jthangam.scs@velsuniv.ac.in*

***Abstract*—Nowadays, cyberbullying is a serious issue that many businesses must deal with. Existing technology makes use of machine learning to automatically detect cyberbullying. Deep learning-based algorithms have been shown to achieve higher accuracy in text classification than existing methods. In this paper, we develop a cyberbullying tweets detection using machine learning algorithm. The proposed method reads the tweets and then classifies the texts relating to cyberbullying and blocks the users. The study uses a k-NN classifier integrated with Deep Learning and show how effective the model is over large text datasets than other methods. The results of simulation shows that the proposed method has higher rate of classification accuracy than the other existing methods.**

***Keywords: Cyberbullying, Detection, Tweets, k-NN Classifier***

## I. Introduction

Significant progress has been achieved in recognising and stopping cyberbullying as a result of the many pieces of research work that has been done in this sector employing a wide array of machine learning and deep learning methodologies. These developments are a direct result of the use of these methods. Very few studies have made use of non-English data for training and assessment purposes. Bangla, Arabic, and Urdu are all instances of such languages. Since there has not been much done to prevent cyberbullying in countries like India, where many people use both English and Hindi language written in Devanagari script, we propose moving forward with our proposed learning algorithm to integrate this data and detect cyberbullying in real-time tweets [1].

Inconsistencies in the classification labels mean that we ca not combine the three datasets in the way they were originally collected. Before moving on with these datasets, we need to settle on a single classification method. For this reason, we opted for a 0–1 classifier that provides instant feedback on whether or not a given text contains data relevant to cyberbullying. Before the data can be organized, it must be cleaned of symbols, URLs, emails, stopwords, white space, numbers, punctuation, stemming, and single tokens [2].

Word embedding is a technique used in natural language processing in which a vector of real numbers is associated with each word or sentence in a lexicon. When training on context-based data, these word embeddings become useful since they reveal hidden relationships between words. Contextual data is required for this kind of instruction. Word embeddings are preferred over other methods because they may be constructed from huge publicly available corpora and do not necessitate expensive annotation. Once this is done, pre-trained embeddings can be used instead of training data for tasks that require only a small amount of labelled data [3].

A key component of the word embeddings library is stacked embeddings. A method called stack embedding is used to merge numerous embeddings into one. When combining regular and contextual string embeddings in the same document, for instance, the stack embedding approach is used. Combinatorial flexibility is made possible via stack embeddings, and it has been proven that certain combinations of embeddings yield the greatest results [4]. Any kind of information can be stored in a stack embedding.

Deep learning (DL), a subfield of machine learning, can be used in various contexts. More neural networks are used in deep learning models, making them more effective than machine learning (ML) models and statistical methods. Therefore, deep learning models are highly favoured in our practise. Text categorization is an area in which deep learning algorithms have been shown to excel, with state-of-the-art results being achieved on a wide variety of traditional academic benchmark problems. The fact that these algorithms have achieved cutting-edge performance demonstrates this. More data is directly correlated with better model accuracy when using deep learning networks [5]. Since hybrid techniques have been found to be excellent models for minimising sentiment mistakes on more precise training data, we compared a hybrid model to a single-layer model. Hybrid methods were used because they have been proven to work.

Hybrid models, especially those that integrated deep learning methodologies, outperformed solo models in detecting cyberbullying on all datasets. In this research, we examine how our proposed hybrid model handles forms written in different languages. In this study, we looked into whether or not integrating many models would be beneficial, and we found that it would be. We looked into how model improvements in trait extraction, node storage, and text classification might be related [6].

In this paper, we develop a cyberbullying tweets detection using machine learning algorithm. The study uses a simple k-NN classifier and show how effective the model is over large text datasets thanother methods. The proposed method reads the tweets and then classifies the texts relating to cyberbullying and blocks the users.

The novelty of the work involves the following:

The k-NN model with increased layers of artificial neural network (ANN), where it is referred as deep learning is applied stacked word embeddings to tweets posted in a wide variety of languages. Our findings suggest that combining the models improves the precision with which cyberbullying can be identified. Once the model has been trained, we'll use Python and the web technology stack to design an interface inspired by Twitter. The information typed into this interface will happen in real time, so it will be able to decide if the content being typed in fits the definition of cyberbullying or not.

## II. Related Work

There are more than a billion people using messaging apps like WhatsApp and Facebook Messenger, and they send and receive more than 30 billion messages per day. Unfortunately, not all text messages are positive because of the abundance of negative content conveyed in personal, social, and professional contexts. Abusive language includes, but is not limited to, words and sentences that insult, display hate speech toward a person or community, threaten a person, or project obscene beliefs. Words or sentences that are threatening or that portray obscene attitudes are other examples of abusive language. More and more people are using software that prevents them from viewing or sharing such materials [7].

Neural network-based models have been used by researchers across several fields to tackle real-time deployment and massive data processing difficulties. For token-based classification, a simple convolutional neural network (CNN) model is applied to pre-trained word vectors. Convolutional neural network is an abbreviation for this. Fifty million tweets were combed through using the CNN algorithm to discover instances of cyberbullying [8]. With the help of a CNN based on pronunciation, cyberbullying like misspelt emails and posts on social media can be avoided. Due to their capacity to recognise spoken language, CNNs are commonly utilised in the process of constructing models for the identification of bullying tweets. As a means of identifying such languages, a Bidirectional Encoder Representations from Transformers (BERT) model was trained using the Reddit Abusive Language Dataset. Indirect speech recognition and natural language processing have already been shown to be effective in spotting cyberbullying SMS communications. Multiple alternative methods exist now for cyberbullying detection, with most of them based on textual and user characteristics. Utilizing a sentence dictionary, we were able to create a simple algorithm for detecting cyberbullying in chat transcripts [9].

Much less frequently, machine learning-based ensemble models designed to spot cyberbullying incidents have been published. Numerous deep learning architectures, including regularization-focused, convolutional, and bidirectional network topologies, were used in the experiments. Experiments were conducted using a set of criteria the authors created for deciding the success of these various architectural approaches. It has been noticed that LSTM models are provided with a larger number of parameters for the purpose of training, while GRU has been found to obtain the best median accuracy with the fewest metrics. One major flaw of hybrid models is that they use the same dataset for both their experiments and their assessments. Automating the removal of inappropriate interactions, like those that are offensive or use bad language, is hard and takes time. Because of the prevalence of racist and sexist language in YouTube comments that encourage controversial debate and are centred on cyberbullying, we made sure to account for these sensitive subjects when building the detection model. Players in bullying situations can be viewed from three perspectives: as the bully, the victim, or the accuser. When looking for patterns of cyberbullying in textual representations, feature engineering is a frequent technique [10]-[12].

In addition, systems with limited capabilities, like those found in on-board vehicle computers and mobile devices, have a hard time processing massive database. The feature extraction and feature selection procedures become more difficult as more attributes are introduced. Research studies are being undertaken to identify cyberbullying, but they do not yet take into account the semantics of words. A Twitter-based model can extract properties from tweets and use them to classify tweets into various categories. Data domain expertise is also put to use in the creation of novel features with the aim of enhancing classifier efficacy. Including factors like the number of pronouns, the length of comments, and the usage of capitalization and emoticons can increase the efficiency of Support Vector Machines (SVM) in recognising cases of cyberbullying. Both n-gram and skip-gram representations share the fact that pronouns are used frequently [4].

Several models uses both textual and visual data from social networks to identify instances of cyberbullying. Researchers have used features of social network analysis (SNA) to group cyberbullying behaviors. These features can be seen as a graph with nodes and edges. Improved cyberbullying detection was achieved by the combination of literary features and the word frequency-inverse document frequency technique. In addition, the SNA purview has expanded to include race, sexual orientation, and intelligence level as tags. Cyberbullying will be easier to find if it is put under more vague and hidden headings [1].

Some of the previously listed research shows that periodic training of a transformer can be very resource costly. Predictions with a high degree of accuracy and precision can be made by combining these models with several types of recurrent neural networks. A novel approach has been created to better identify malicious text in transit over a network and limit access to it. While consuming extremely little computational resources and processing data at a steady rate, ML is able to recognise abusive text with a high level of accuracy at the edge of the network and in the cloud.

## III. Proposed Method

It has been proposed to use social media as a proof-of-concept for an automated cyberbullying monitoring and management system. Data cleaning and various forms of pre-processing come first, before the training data is fed into stacked word embeddings. Our ultimate goal is to train the KNN model to achieve results superior to those achieved by individually taught models of deep learning.

The model will be archived for later use in the finalised version of the product. The user can choose from a wide variety of features. The site operates similarly to social networking platforms. The content status can only be seen by the administrators. While this is still very much a prototype, it does show promising early signs of development toward a more polished final product.

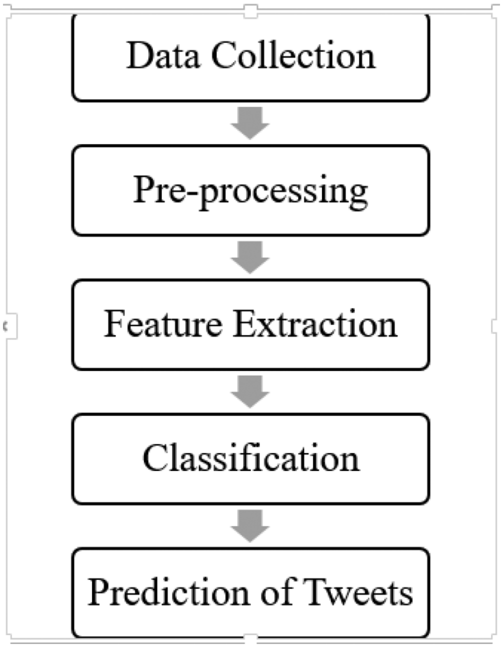

Fig. 1: Proposed Model

### A. Data Collection

The whole communication history of the user is saved in a plaintext file every three hours, with a newline character between each utterance. Every three hours, this file gets updated. After collecting the data, it is cleaned by stripping each sentence of its HTML and characters other than the alphabet. Once the sentences have been cleaned, they are tokenized into a word list and then one-hot encoded into binary vectors. These vectors are the raw data that the KNN model uses to make predictions.

KNN analyses each text for the presence of objectionable language by processing the input vectors. Once the entities have been labelled, they will be used to start a training dataset for the deep learning model. Using this data set for training, the model classification metrics will be improved, and the weight matrix will be updated. Abusive sentences are found, highlighted, and labelled according to what kind of abuse was found.

If a KNN probability of predicting the presence of an abusive class of text in the dataset is greater than 0.8, then it is added to the dataset for training purposes. To ensure that only clean, correctly labelled data is used in the future, we have set the threshold for obtaining training data at 0.8. To make room for the next iteration of the textual data collection process after the KNN model has been trained, these entities are deleted entirely from the MaLang app. On average, CASE needs 15 minutes to complete data cleaning, tokenization, and training.

### B. Classification

The dataset is taken into account to help detect possibly harmful text. The dataset includes over 28,000 sentences that have been labelled as either non-offensive, offensive, or identity-hate speech. We divided the dataset into two categories: abusive and non-abusive material. Twenty-five thousand sentences were cleaned up by converting them to lowercase and removing non-alphabetic letters and HTML tags. The evaluating sample size was three thousand sentences. When everything was spelled in lower case, both sets were put to use.

To ensure uniformity in the output, we first used the NLTK software to tokenize the text and then clipped each sentence to the same length. After that, we used these to build an entity array of terms for each sentence. At the same time, a one-hot encoding was carried out on the vocabulary of each item, resulting in the generation of a matrix in which each class was represented by a binary vector. The vectors are fed into the embedding layer for the CASE deep learning model.

Each participant will emerge from the initiative with a singular identity. The label is a random five-digit string that can be alphabetic or numeric. The label may consist of letters, numbers, or both. No more than one person may wear this tag at any given time. After entities are sorted into those that are abusive and those that are not, the latter are associated with the user ID and delivered to a cloud-based module for further processing. No matter how many devices or users contributed to the data, the cloud-based service gets all tagged objects that match the user ID.

### *C. Training using DL*

The models are trained using the deep layers of ANN, where the model is trained using the input parameters. The study uses sigmoid loss function with cross-entropy to reduce the losses existing between the layers:

$F(s) = (1+e^{-s})^{-1}$

$\mathrm{CE}=-t_1\log(f(s_1))– (1-t_1)\log(1-f(s_1))$

Upon training from the input training datasets, the study uses the results for classifying the instances of the test dataset using KNN-Classification.

### *D. KNN-Classification*

k-Nearest Neighbor (K-NN), is a popular classifier that uses a modification of the nearest neighbour technique. In this variant of the algorithm, rather than relying on the vote of a single nearest neighbour, the classification of a mystery sample is based on the votes of k such neighbours.

K-Nearest Neighbor has been shown to be an effective nonparametric supervised pattern classifier that is also straightforward to implement. Using this method, things are put into groups based on how much they look like their training examples in the feature space. An easy way to use the k-nearest neighbour method is to put an object in the same category as the majority of its k closest neighbors.

The training samples typically have vector labels in a feature space. K-Nearest Neighbors (K-NN) simply requires the feature vectors and the class labels to be stored throughout the training phase. The unlabeled vector (also called a query or test point) is classified by assigning it the label that appears most frequently in the k training examples that are physically nearest to it.

### *E. Algorithm*

**Step 1:** Calculate the distance between each input record and each training record.

**Step 2:** K-nearest neighbour is chosen after the training records are sorted by distance.

**Step 3:** Our neighbours are through tapping into the majority-owning class.

The approach uses the most common class among the input record neighbours to identify the input record class.

## IV. Results And Discussion

Multiple perspectives on the feasibility of the proposed paradigm are examined here. When applied to various forms of training data and quantities of residuals, how well do the feature selection methods perform in terms of accuracy of classification. Using 60% of the available training data, we looked at how well different methods of selecting features classified the data.

Over 98% of information can be classified by the system-level module into abusive and non-abusive text messages for smartphones or personal computers. Only potentially harmful content, in the form of entities that define the text via a collection of tokens, is then uploaded to the cloud. This takes place once the information has been sorted into relevant categories. With the exception of tokenized text messages, the user data is never sent to the cloud, so their privacy is protected.

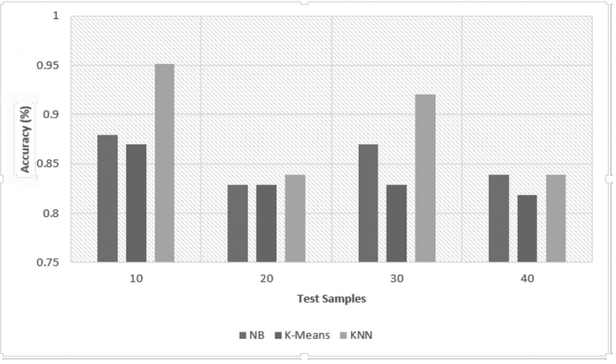

Fig. 2: Accuracy

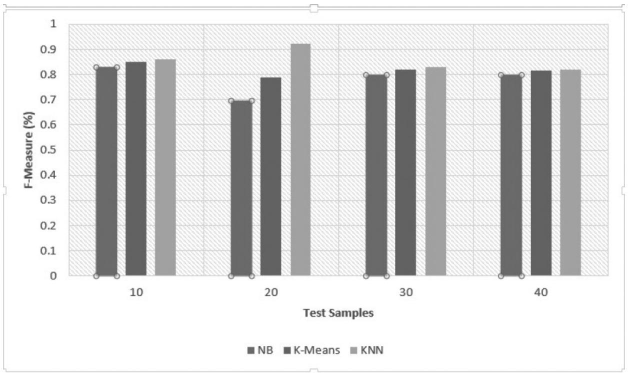

Fig. 3: F-Measure

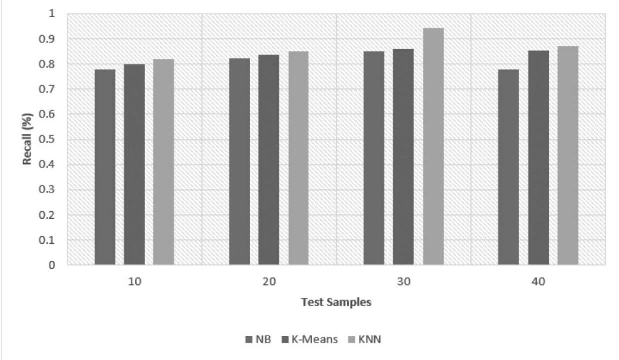

Fig. 4: Recall

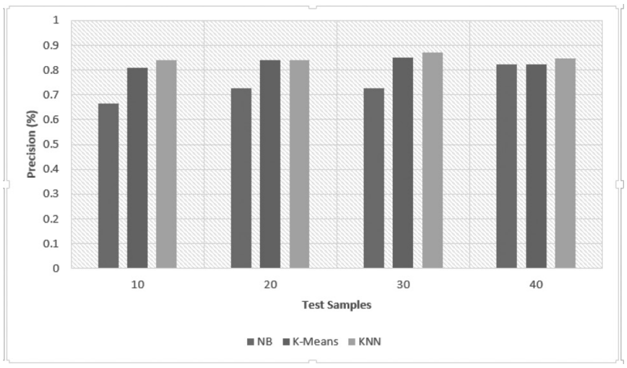

Fig. 5: Precision

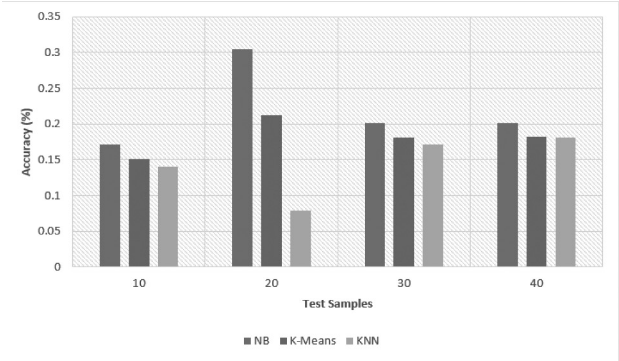

Fig. 6: Classification Error

Because context clues like articles and stop words are left out of tokenized data stored in the cloud, a person cannot understand it. With only their knowledge and experience, it is very unlikely that a person could put together the bits of data that have been sent to the cloud into a coherent message.

Due to the fact that 88% of the classification occurs on the user device, the cloud-level module requires less data storage and processing time to determine if a text is dangerous. It is expected that using KNN will reduce the time and resources that cyberbullying detection systems need by 38% to 42%.

During KNN training, progress was made at a rate of 0.001 seconds per minute. The Adam optimizer was utilised, along with a categorical cross-entropy loss function. The model had an average accuracy of 85%. In general, KNN was able to achieve a 90% accurate prediction rate and a 91% average recall rate. Keeping employee information secret is one of the main benefits of the decentralised approach. Additionally, the decentralised approach allows this technology to scale to extremely extensive business data networks.

Further, the comparison with the training and test datasets is provided in the paper. It is seen that after training, the testing using KNN classification achieves higher rate of accuracy than the other methods. The results of comparison over various cyberbullying datasets is provided in Table 1.

Table 1: Results of Training/Test Dataset

| Dataset | Training | Testing |
|---|---|---|
| 10 | 90.982 | 91.910 |
| 20 | 91.033 | 91.912 |
| 30 | 91.381 | 92.264 |
| 40 | 91.547 | 92.399 |
| 50 | 92.035 | 92.891 |
| 60 | 92.045 | 92.935 |
| 70 | 92.409 | 93.302 |
| 80 | 92.741 | 93.637 |
| 90 | 93.548 | 94.418 |
| 100 | 93.893 | 94.800 |

## V. Conclusion

In this study, we created a machine learning-based method to identify cyberbullying in tweets. The proposed system scans tweets for keywords related to cyberbullying and suspends the accounts of individuals who use them inappropriately. In order to show that the model performs better on huge text datasets, the study employs a simple k-NN classifier. This research discusses the challenge of automatically identifying cyberbullying text in datasets with many languages and proposes a solution. Finding a way to overcome this obstacle is crucial if we ever want to exert some kind of control over the material that is shared between languages and protect users from potentially harmful comments like verbal attacks and coarse language. The results of simulations show that the proposed method is better at classifying data than the best methods available right now.

## References

[1] Zhu,C., Huang,S., Evans, R., &Zhang,W. (2021). Cyberbullying among adolescents and children: a comprehensive review of the global situation, risk factors, and preventive measures, Frontiers in public health, 9, 634909.

[2] Evangelio, C. Rodriguez – Gonzalez, P., Fernandez – Rio, J., & Gonzalez – Villora,S(2022). Cyberbulling in elementary and middle schools students: A systematic review. Computers & Education, 176, 104356.

[3] Giumetti, G.W., & Kowalski, R. M. (2022). Cyberbullying via social media and well being. Current Opinion in Psychology, 101314.

[4] Barlett,C.P., Simmers, M.M., Roth, B., &Gentile,D. (2021). Comparing Cyberbullying prevalence and process before and during the COVID – 19 Pandemic. The Journal of Social Psychology, 161 (4), 48 – 418.

[5] Yang, F..(2021). Coping Strategies, cyberbullingbehaviors, and depression among Chinese netizens during the COVID - 19 pandemic: a web based Nation Wide Survey. Journal of affective disorders, 281, 138-144.

[6] Doral,O., &Mishara, B. L . (2021). Systematic review of risk and protective factors for suicidal and self-harm behaviors among children and adolescents involved with cyberbulling. Preventive medicine, 152,106684.

[7] Chan, T.K., Cheung, C. M., & Lee, Z.W (2021). Cyberbullying on social networking sites: A literature review and future research directions. Information & Management, 58(2), 103411.

[8] Leduc, K., Nagar, P.M Caivano, O., & Talwar, V. (2022).” The thing is, it follows you everywhere”. Child and adolescent conceptions of cyberbullying. Computers in Human Behavior, 130, 107180.

[9] Yokotani, K., & Takano, M. (2021) .Social contagion of cyberbulling via online perpetrator and victim networks. Computers in human behavior, 119, 106719.

[10] Eyuboglu, M., Eyubogly,D., Pala, S.C., Oktar,D., Demirtas, Z., Arslantas,D., &Unsal, A. (2021). Traditional school bullying and cyberbullying: Prevalence, the effect on mental health problems and self -harm behavior. Pyschiatry research, 297, 113730.

[11] Yuvaraj, N., Srihari,K., Dhiman, G., SOmasundaram, K., Sharma, A., Rajeskannan, S.M.G.S M.A., …. &Masud, M (2021). Nature – inspired – based approach for automated cyberbullying classification on multimedia social networking Mathematical Problems in Engineering, 2021.

[12] Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., &Rajan, A. R. (2021). Automatic detection of cyberbullying using multi – feature based artificial intelligence with deep decision tree classification. Computers & Electrical Engineering, 92, 107186.