Proceedings of the Sixth International Conference on Intelligent Computing and Control Systems (ICICCS 2022)
IEEE Xplore Part Number: CFP22K74-ART; ISBN: 978-1-6654-1035-9

2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1035-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCS53718.2022.9788299

A Study on Automatic Detection of Cyberbullying using Machine Learning

Asif Ahmad Khan
(Student)
asifahmadkhan_2k20cse07@dtu.ac.in
Delhi Technological University
Delhi – 110042, India.

Dr Aruna Bhat
(Associate Professor)
aruna.bhat@dtu.ac.in
Delhi Technological University
Delhi – 110042, India.

***Abstract*- With time, use of the internet has become very common among people, and thus the rise of internet usage has given birth to a problem of cyberbullying. Cyberbullying can have a serious impact on the psychological health of the person who is the victim of it. Hence, detection of cyberbullying is required on the internet or social media. Much research has been done in the field of detection of cyberbullying. Machine learning can be the one of the approaches that is used for automatic cyberbullying detection. This paper has studied some of the papers related to cyberbullying. Moreover, some NLP techniques and different models used for cyberbullying detection tasks have been reviewed. The graph, which is based on the papers reviewed, shows that the tf-idf is mostly used either directly or with a combination of other techniques for feature extraction in cyberbullying detection using machine learning.**

***Keywords* - Cyberbullying, Natural Language Processing, Machine Learning, Featurization.**

I. Introduction

Nowadays, the internet is easily accessible to all and most of us spend a large amount of time on the internet through online education, online gamings, e-commerce, social networking sites etc. With the increase of internet usage, the problem of cyberbullying also arises. Cyberbullying can be detrimental to a person's psychological health which can even make someone to commit suicide. Hence, monitoring is required for cyberbullying detection on the Internet/Social media. Many works have been done for the automatic cyberbullying detection using different approaches. One of the methods for the automatic cyberbullying detection is by using supervised machine learning techniques. Patterns used by bullies in their languages should be detected for cyberbullying detection tasks, and ML learning can be useful for this pattern detection [11]. In this supervised learning technique, we will have a dataset that is already labelled as bully or non_bully text, which will be further used for training various classification models. Once the training of an ML model with a given dataset is completed, the model is ready to predict the result of given text.

In the automatic detection of cyberbullying, binary classification is done on text to classify whether it is bully text or not. Multiclass classification is also performed in cyberbullying detection., Talpur, B.A., et al [10] performed multi class classification based on severity of cyberbullying . A dataset comprising a set of text data can not be directly used for a classification task. First, we need to convert this text into an n-dimensional input vector and this input vector can be given to different models for the classification of text. This text can be converted into input vectors using various NLP (Natural language Processing) techniques like

BOW, TF-IDF etc . We have also reviewed some of the NLP techniques used for cyberbullying detection. To develop a successful ML model for detecting cyberbullying behaviour, various factors play a role, and the features used in the task of cyberbullying detection are the most important factor [10]. In order to make our ML model recognize and classify our text for the cyberbullying detection task, text is preprocessed and useful information is analysed from it [1].

II. Literature Review

Alotaibi, M., et. al [1] proposed the automatic cyberbullying detection method by combining the features of 3 different models of deep learning and these 3 models are transformer block, CNN, and BiGRU. The proposed method classifies Twitter comments as offensive or not offensive. The authors combined the three known datasets and then the performance of the proposed method was measured. Seventy five percent of the data is selected as training data and the remaining data is selected as test data. The proposed model gives an accuracy of 87.99%. Apart from accuracy, the proposed method is also evaluated on four other different performance metrics.

Khairy, M. , et al [2] presented a survey work on the Arabic content for the automatic cyberbullying detection and abusive language. Authors analyse 27 studies on the contents which are in Arabic , among which 10 are on cyberbullying detection and 17 are on detection of the language which are offensive. In contrast to the definition of cyberbullying as a recurrent behaviour, all of the datasets used in the cyberbullying detection method were labelled with a one post. Most of the datasets used are imbalanced, which has an impact on the classifier's performance.

Rosa, H. , et al [3] did a detailed review of twenty two studies on automatic detection of cyberbullying and an experiment to validate current practices using feature engineering and two datasets. The authors use a quantitative systematic review approach for the automatic detection of cyberbullying.

Mozafari, M., et al [4] introduce a novel approach of transfer learning based on BERT to improve the detection of hate speech system's performance. BERT is an existing pre-trained language model . This model addresses the issue such as insufficient amount of labelled data of hate speech. 2 datasets which are available publicly are used by authors. Some biases while collecting the dataset can be detected by the model is observed in the result.

Isaac, Akileng,et al[5] uses knowledge distillation to enhance the model of LSTM which is used for hate speech detection.

Al-Ajlan, M.A. , et. al [6] proposed a novelty algorithm CNN-CB for cyberbullying detection using convolutional neural network. Proposed algorithm does not require feature engineering in detection of Cyberbullying. Algorithm uses word embedding concept and it is performing better than traditional approaches for Cyberbullying detection task.
Monika, et al [7] worked on big data associated with the crime using Apache pig.

Perera, A. , et al [8] presented a cyberbullying detection solution to find cyberbullying precisely along with the themes/categories related to cyberbullying using natural language processing (NLP) and supervised Machine Learning. Logistic Regression and SVM classifier is used in this cyberbullying detection system. Besides Tf/Idf, n gram and profanity along with sentiment analysis improves the system's accuracy. Accuracy of the solution

proposed by Authors is 74.50%. Sarcasm text is not detected as cyberbullying in this proposed solution.

Shah, R., et al [9] presented a cyberbullying detection system on the Twitter dataset. The distribution of the dataset as non bully and bully text is equal. The Tf-idf method is used for feature extraction. Different classification models are used for cyberbullying detection. As a result, the authors found that Logistic Regression performed best among all classification models, with an accuracy of 93% and precision of 91%.

Talpur, B.A., et al [10] developed a cyberbullying detection feature-based machine learning model. Authors introduced a technique to create a new input feature. Along with the new input feature, other predicted features and Twitter API features were used. The results of the model are classified into multiple classes of non-cyberbullied, low, medium, and high levels. The model gives an accuracy of 93%.

Hani , J., et al [11] proposed an approach for cyberbullying detection using ML. For extracting the feature they have used sentiment analysis algorithm and tf-idf method. Classification tasks are evaluated using NN (neural network) and SVM classifiers on different n gram language models. Neural networks perform better than SVM. The accuracy of SVM with 4-gram is 90.3% and the accuracy of a neural network with 3-gram is 92.8%. The size of training data is limited for detecting patterns in cyberbullying, so to further enhance the performance of the model, a larger data size is required.

M. Ahmed, et al [12] performed sentiment analysis on data from Twitter. Authors use 3 ML models to classify the sentiment of a tweet into 5 categories.

Muneer,A., et al[13] has performed a comparative study of the model for cyberbullying detection with a global dataset compiled with unique tweets from Twitter. Performance is compared using seven machine learning models. Authors observe that performance of logistic regression classifiers improve with the increase of data size .For extracting the feature, tf/idf and word2vec are used. Among the seven classifiers, logistic regression performed best on the compiled global data set. The F1 score of the LR model is 0.9280, and its accuracy is 90.57%.

Lepe-Faúndez, M., et al [14] proposed different models using hybrid approaches that combine lexicons and machine learning for detection of aggressiveness in Spanish language. Five distinct ways to construct different models are proposed, each with its own way of extracting features from text. As a result, a hybrid model that uses lexicons provides the best results in the 3 language corpora of Spanish when compared to a model which does not use lexicons.

Ali, W. N. H. W., et al[15] proposed a model based on machine learning for cyberbullying detection using techniques like hyperparameter optimization , resampling and feature selection. SVC Linear and Decision Tree are used .Word-n grams technique is used for feature extraction. Eight various experiment setting were done to test the classifier , experiment setting like classifier + smote +feature selection , classifier + hyperparameter optimization etc. When tested using the x square test (feature selection) without any use of hyperparameter optimization and resampling, the Decision Tree classifier outperforms the SVC Linear classifier is shown in the result.

AlHarbi, B. Y., et al [17] presented a lexicon based approach for cyberbullying detection using sentiment analysis. PMI, Entropy, and

Chi-square are the three different lexicon approaches used. A comparison is made among these three lexicon approaches to find which one is better for cyberbullying detection in Arabic text. Among all 3 lexicon approaches, the PMI approach gives the best results as compared to the remaining two approaches for cyberbullying detection in Arabic.

III. Methodology for Cyberbullying Detection using supervised Machine Learning

General steps that are followed for automatic Cyberbullying Detection using supervised machine learning are shown in fig1. Each of these steps are explained below.

Data Collection and Data Preprocessing-

Data is collected from social media or the internet for cyberbullying detection tasks, then data is preprocessed like handling missing values, removing stopwords or any special symbols, stemming, etc.

We can remove stopwords from text by comparing each word of text to a list of stopwords if a word contains in a list of stopwords, it will be removed from text. Python nltk library can be used to get a list of stopwords.

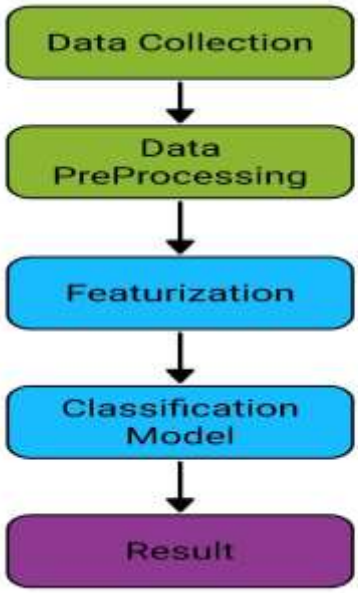

Fig 1. Diagram for general steps of cyberbullying detection.

Featurization-

After data preprocessing, features are extracted from text and various feature extraction methods are used for converting text into n-dimensional vectors. If the two texts are similar or closer to each other, then their vectors should also be geometrically closer to each other, and if the texts are not closer or dissimilar, then their vectors should not be geometrically closer to each other. Such properties should be posed by n-dimensional input vectors. We can also use functions from sklearn python library to implement BoW(Bag of Words) and the Tf-Idf technique to convert text into n-dimensional vectors. One issue with these two techniques is that the n-dimensional vectors created by these techniques are sparse in nature.

Various techniques used for feature extraction in cyberbullying detection tasks,

Talpur, B.A., et al [10] introduced a new input feature and combined it with twitter API features and predicted features (gender,age etc) to create a n dimensional input vector. Hani , J., et al [11] created an input vector by adding sentiment analysis feature with tf-idf, Muneer, A., et al[13] uses tf-idf and word2Vec to get n dimensional

input vector from text. Shah, R., et al [9] uses tf-idf technique for the feature extraction. Lepe-Faúndez, M., et al [14] uses various ways to create input vectors for model using lexicon approach , combining lexicon with tf-idf, combining lexicon with word embedding , combining lexicon with tf-idf and word embedding . Ali, W. N. H. W., et al[15] uses n-grams bag of words technique for getting features from text.

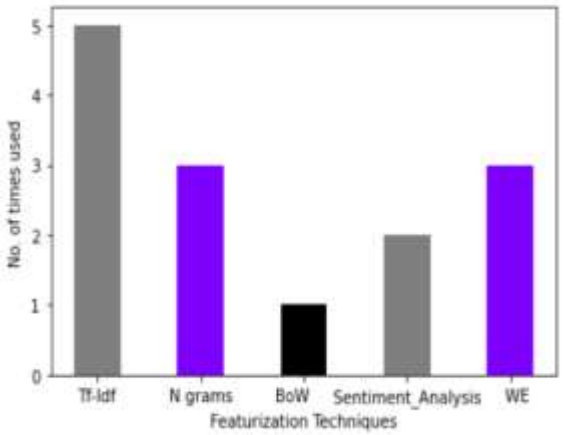

Fig 2. Graph for estimated no. of times featurization techniques used

The estimated numbers of some of the techniques used for featurization in the cyberbullying detection task based on the papers we reviewed are shown graphically in fig 2. The graph shows that tf- idf is used more frequently as compared to other techniques followed by WE(Word Embeddings) and N-grams in cyberbullying detection.

Classification Model-

The n-dimensional vectors are now given as input for training of classification models. Different ML models and DL models can be used for the classification of cyberbullying texts, even combinations of different models can be used for classification tasks.

Classification models used in some of the study of cyberbullying detection are Ali, W. N. H. W., et al[15] uses SVC Linear and Decision Tree model, Talpur, B.A., et al [10] uses Naive Bayes , SVM with RBF Kernel, Random Forest, and KNN model , Shah, R., et al [9] uses Multinomial NB, Random Forest, Logistic Regression, SVC, and SGD model , Hani , J., et al [11] uses Neural network and SVM model.

The most crucial step in the architecture of text classification is choosing the best classifiers[10]. Various performance metrics such as accuracy, confusion matrix, log loss, F1-score, AUC, etc can be used to compare the effectiveness of different models for the automatic detection of cyberbullying tasks for a given dataset and can help to choose the best model.

Result/Output-

Once the models are trained, they are used to classify our input text as whether it is bullying text or not.

IV. Limitation of using traditional ML for Cyberbullying detection

Approaches based on traditional machine learning algorithms rely largely on handmade features that are generally time-consuming and incomplete [1]

Table 1. Various methods and models used for cyberbullying detection

	Authors	Model	Method Used	Result	Language
1.	Alotaibi, M., et. al [1]	Transformer block , CNN and BiGRU.	Combining features of 3 DL models.	Accuracy -87.99%	English
2.	Hani , J., et al [11]	SVM and Neural Network	Sentiment Analysis , N-grams and Tf-IDF method.	Accuracy (SVM) - 90.3% and Accuracy (NN) - 92.8%	English
3.	Talpur, B.A., et al [10]	NB, SVM with RBF Kernel, Random Forest, KNN.	New input feature , other predicted features and Twitter api features.	Accuracy -93%	English
4.	Shah, R., et al [9]	SVC, Multinomial NB, Logistic Regression, RF and SGD.	TF-IDF.	Accuracy (LR) -93%	English
5.	Lepe-Faúndez, M., et al[14]	22 different Model	5 Approaches for creating model Lexicon, WE_Lexicon, TF_IDF_Lexicon, WE_Lexicon_TF-IDF, and Ensemble approach.	Accuracies- Mexican corpus - 0.8431 Chilean corpus - 0.892 Chilean-Mexican corpus- 0.8548	Spanish
6.	Muneer,A., et.al[13]	Seven Different ML Classifiers	TF-IDF and Word2Vec.	Accuracy (Logistic Regression) -90.57%	English
7.	Perera, A. , et al [8]	SVM and Logistic Regression	TF-IDF , N-gram, profanity, and sentiment analysis.	Accuracy -74.50%	English
8.	Al-Ajlan,M.A., et. al [6]	CNN-CB	Word Embedding.	Accuracy -95%	English
9.	AlHarbi, B. Y., et al [17]	—	Sentiment Analysis and Lexicon Approaches	Avg. F-score for PMI -81%	Arabic

			PMI, Entropy and Chi-square.		
10.	Ali, W. N. H. W., et al [15]	SVC Linear and Decision Tree	Techniques hyperparameter optimization , resampling and feature selection. N-grams BoW.	Accuracy using default parameter (SVC Linear) -95.54%	English

V. Conclusion

As we are all aware, the problem of cyberbullying over the use of the internet in our time is experienced by many internet/social media users. Cyberbullying is undesirable as it can have a bad impact on the psychological health of those who are bullied. Therefore, an automatic system for detection of cyberbullying should be implemented over the internet to control the problem of cyberbullying. Work has been done in the area of automatic detection of cyberbullying with the use of different feature extraction techniques, different ML and DL models, and even combinations of models.We have studied some papers in the area of cyberbullying detection. We have also reviewed different models and different methods used for cyberbullying detection using machine learning. Tf- idf technique is mostly used with or without using another technique for getting input vectors from text in the task of cyberbullying detection, which uses an ML approach. We have also found in some studies of cyberbullying detection that logistic regression classifiers perform better when compared to other ML classifiers.

In our future work we would like to explore deep learning models for cyberbullying detection.

References

[1] Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, *10*(21), 2664.

[2] Khairy, M., Mahmoud, T. M., & Abd-El-Hafeez, T. (2021). Automatic Detection of Cyberbullying and Abusive Language in Arabic Content on Social Networks: A Survey. Procedia Computer Science, 189, 156-166.

[3] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.

[4] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019, December). A BERT-based transfer learning approach for hate speech detection in online social media. In International Conference on Complex Networks and Their Applications (pp. 928-940). Springer, Cham.

[5] Isaac, Akileng & Bhat, Aruna. (2022). A Conceptual Enhancement of LSTM Using Knowledge Distillation for Hate Speech Detection. 10.1007/978-981-16-4016-2_53.

[6] Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep learning algorithm for cyberbullying detection. International Journal of Advanced Computer Science and Applications, 9(9), 199-205.

[7] Monika and A. Bhat, "An analysis of Crime data under Apache Pig on Big Data," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2019, pp. 330-335, doi: 10.1109/I-SMAC47947.2019.9032565.

[8] Perera, A., & Fernando, P. (2021). Accurate Cyberbullying Detection and Prevention on Social Media. Procedia Computer Science, 181, 605-611.

[9] Shah, R., Aparajit, S., Chopdekar, R., & Patil, R. (2020). Machine Learning based Approach for Detection of Cyberbullying Tweets. Int. J. Comput. Appl, 175(37), 51-56.

[10] Talpur, B. A., & O'Sullivan, D. (2020, December). Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyberbullying in Twitter. In Informatics (Vol. 7, No. 4, p. 52). Multidisciplinary Digital Publishing Institute.

[11] Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. International Journal of Advanced Computer Science and Applications, 10(5), 703-707.

[12] M. Ahmed, M. Goel, R. Kumar and A. Bhat, "Sentiment Analysis on Twitter using Ordinal Regression," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), 2021, pp. 1-4, doi: 10.1109/SMARTGENCON51891.2021.9645751.

[13] Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet, 12(11), 187.

[14] Lepe-Faúndez, M., Segura-Navarrete, A., Vidal-Castro, C., Martínez-Araneda, C., & Rubio-Manzano, C. (2021). Detecting Aggressiveness in Tweets: A Hybrid Model for Detecting Cyberbullying in the Spanish Language. Applied Sciences, 11(22), 10706.

[15] Ali, W. N. H. W., Mohd, M., & Fauzi, F. (2018, November).

Cyberbullying detection: an overview. In 2018 Cyber Resilience Conference (CRC) (pp. 1-3). IEEE.

[16] R. Kumar and A. Bhat, "An Analysis On Sarcasm Detection Over Twitter During COVID-19," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-6, doi: 10.1109/INCET51464.2021.9456392.

[17] AlHarbi, B. Y., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M., Alshobaili, J. F., & Ibrahim, D. M. (2019). Automatic cyber bullying detection in Arabic social media. Int. J. Eng. Res. Technol, 12(12), 2330-2335.

[18] Garg A., Aggarwal K., Saxena M., Bhat A. (2021) Classifying Medical Histology Images Using Computationally Efficient CNNs Through Distilling Knowledge. In: Tavares J.M.R.S., Chakrabarti S., Bhattacharya A., Ghatak S. (eds) Emerging Technologies in Data Mining and Information Security. Lecture Notes in Networks and Systems, vol 164. Springer, Singapore. https://doi.org/10.1007/978-981-15-9774-9_66