Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021).
IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4

2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) | 978-1-6654-1960-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICICV50876.2021.9388499

Cyberbullying Detection: An Ensemble Based Machine Learning Approach

Kazi Saeed Alam
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh
saeed.alam@cse.kuet.ac.bd

Shovan Bhowmik
Department of Computer Science and Engineering
Bangladesh Army International University of Science and Technology
Cumilla Cantonment, Bangladesh
bhowmik.sshovon5795@gmail.com

Priyo Ranjan Kundu Prosun
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh
priyo.prosun1997@gmail.com

***Abstract*—Research on cyberbullying detection is gaining increasing attention in recent years as both individual victims and societies are greatly affected by it. Moreover, ease of access to social media platforms such as Facebook, Instagram, Twitter, etc. has led to an exponential increase in the mistreatment of people in the form of hateful messages, bullying, sexism, racism, aggressive content, harassment, toxic comment etc. Thus there is an extensive need to identify, control and reduce the bullying contents spread over social media sites, which has motivated us to conduct this research to automate the detection process of offensive language or cyberbullying. Our main aim is to build single and double ensemble-based voting model to classify the contents into two groups: 'offensive' or 'non-offensive'. For this purpose, we have chosen four machine learning classifiers and three ensemble models with two different feature extraction techniques combined with various n-gram analysis on a dataset extracted from Twitter. In our work, Logistic Regression and Bagging ensemble model classifier have performed individually best in detecting cyberbullying which has been outperformed by our proposed SLE and DLE voting classifiers. Our proposed SLE and DLE models yield the best performance of 96% when TF-IDF (Unigram) feature extraction is applied with K-Fold cross-validation.**

***Index Terms*—Offensive Language, Cyberbullying, Text Classification, Machine Learning Classifiers, Twitter.**

I. Introduction

The increasing use of online platforms has created a powerful influence over people which enables them to express their views and ideas freely than ever before. Social media sites such as Twitter and Facebook has become an integral part of our day-to-day life due to the enormous popularity among people, particularly among teenagers. Certainly, the substantial increase in the usage of these platforms has also some negative consequences as well as these teenagers are often exposed to various behavioral and psychological threats. Cyberbullying in the form of influential social attacks is one of the potential sources of these threats. Moreover, social media platforms allow people to harness the option of being anonymous and hiding their self-identity, which can lead to misusing the technical features to satisfy their unkind deeds. Things also become worse when bullying occurs more frequently over time.

Offensive language containing abusive behavior with the purpose of creating harm to others has become the most dangerous activity on social networking platforms. Cyberbullying can take various forms such as aggressive content, harassment, toxic comment, hate speech, sexism, racism etc. [1] Most of the time these hateful texts lead to terrible mental health effects such as anxiety, depression, self-harm, social and emotional perplexity even suicidal thoughts or attempted suicides. [2] A study by the Pew Research Center [3] showed that over 60% of the US people on social media platforms have been exposed to cyberbullying, with teenager particularly girls enduring the worst forms of it. Thus, cyberbullying has turned into a global problem in the form of an epidemic.

Therefore, to alleviate such heinous acts of cyberbullying, many global preventive and intervention approaches have been introduced with the aim of safety improvement of internet users. However, offensive text detection is considered a challenging and complex task even for humans except the victims owing to the variations of the language used. People often use sarcasm, intimidation, coarse language, and colloquialisms in a friendly manner without the intention to harm others. So, it is a task of great significance that is attracting researchers all over the world to work on the identification of whether a post or a tweet is offensive or not.

Inspired by the need for automated detection of offensive language, we work on detecting offensive posts withdrawn from Twitter. We have used several Machine Learning Models and Feature Extraction techniques to implement a Single Level Ensemble Model (SLE) and a Double Level Ensemble Model (DLE) architecture to automatically identify posts and classify them into two groups as 'offensive' and 'non-offensive'. Our novel model architecture has superiority in performance compared to four different classification techniques, namely, Multinomial Naive Bayes (MNB), Logistic Regression (LR), Decision Tree Model (DT), Linear Support Vector Classifier

(LSVC) and three Ensemble strategies known as Gradient Boosting Classifier (GBoost), AdaBoost(AdB) and Bagging. Using a dataset assembled of both offensive and non-offensive posts, We have performed experimental evaluation and obtained very promising results. In addition to that, to verify the quality of the collected data, we have also carried out four cross-validation strategies and report the assessments in three performance metrics (Accuracy, F1-Score and AuC).

II. Related Works

Recently several approaches have been introduced for detecting social media bullying. In this segment, the closely related research work that has been previously done detecting cyberbullying on social media sites is briefly described.

Dalvi et al. proposed a machine learning model in order to identify and prevent cyberbullying on Twitter [4]. For the training and testing of social platform bullying items, two classifier models were used, namely, Support Vector Machine (SVM) [5] and Naive Bayes [6]. Both the classifiers were able to identify the true positive scenarios, maintaining 71.25% and 52.70% accuracy individually. But SVM surpassed Naive Bayes on the same dataset for equivalent work. Rezvani et al. presented an intelligent cyberbullying identification system [7] which: (i) extracts attributes from the image, image metadata and textual items; (ii) contextualizes the drawn out attributes by creating a crowdsourced feedback loop; (iii) merges the attributes with a neural network to classify and develop potentially handy attributes. Their proposed approach has been able to substantially upgrade most metrics with the incorporation of contextual attributes. Talpur et al. developed a feature-based system to detect the magnitude of cyberbullying in a tweet [8]. This architecture utilizes tweet-content features in order to create an ML classifier for identifying tweets as non-cyberbullied, and moderate, medium and extreme bullied tweets. In this analysis, pointwise semantic orientation has been brought in as a new input feature and it provides a promising result with 93% accuracy and 92% F-measure. But this research work restricts the training of artificial neural network models due to strongly biased positive categories.

Muneer et al. proposed an automatic cyberbullying detection system by assembling a universal dataset consisting of more than 35,000 distinctive tweets [9]. In addition, seven different classifier models were employed i.e. Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Light Gradient Boosting Machine (LGBM), AdaBoost (ADB) and Random Forest (RF). The experimental findings indicate the dominance of LR, which obtained a median accuracy of approximately 90.57%, however not many feature extraction methods were investigated in this study. Balakrishnan et al. proposed an automated cyberbullying identification model using a 'Twitter' dataset comprised of 5453 unique tweets [10]. Three ML classifier algorithms namely, Random Forest, Naive Bayes and J48 were applied to categorize each tweet into one of four classes: 'normal', 'bully', 'aggressor' and 'spammer'. Key features were interpreted by the dimensionality reduction technique and merged into a single model that provided the highest detection accuracy. But this study lacked the usage of cross-validation techniques which resulted in an imbalanced dataset. Herath et al. presented an automatic offensive language detection system [11], which took the challenge of solving three different sub-tasks- (A) recognizing offensive language, (B) categorizing the type of offense and (C) identifying offense target. Their developed ensemble models, based on deep learning approaches, pulled off F1-scores of 0.906, 0.552 and 0.623 in those three corresponding sub-tasks.

Lu et al. proposed a Character-level CNN with Shortcuts (Char-CNNS) architecture [12] in order to detect if a social media text/sentence comprises cyberbullying or not. Characters were used as the smallest units of learning, which allowed the architecture to control spelling mistakes and deliberate confusion in real world sentences. Also, a focal loss function was used to resolve the class imbalance challenge. Shah et al. propounded a simple ML based approach for cyberbullying detection [13]. A uniformly distributed twitter dataset was fed to several ML models, of which the logistic regression (LR) classifier produced the most precise categorization of bullying and non-bullying tweets providing 91% precision and 93% F1-score. Although various ML models were used, no ensemble techniques were incorporated. Putri et al. presented a detection system in order to automatically recognize hateful tweets in the Indonesian language [14]. This application employed various ML classification algorithms as well as compared the functioning of the architecture using SMOTE in order to get over imbalanced data. Among all the classifiers, Multinomial Naive Bayes (MNB) produces the finest result proving 93.2% recall and 71.2% accuracy. In short, several current works [15] concentrate on the development of detection models that can identify cyberbullying and hateful texts.

III. Proposed Detection Model

In our proposed model, We have used a publicly accessible dataset in this research to validate our three-level voting scheme with the help of several well established machine learning (ML) algorithms and ensemble techniques. Figure 1 represents the overall working process of our work.

A. Dataset Description

Cyberbullying, which is also known as cyberharassment, has become increasingly common, especially among teenagers as the digital sphere has expanded and technology has advanced. There are various datasets available that contain hateful texts and offensive languages used by different ages of people on various social media websites. As we are concentrating on detecting cyberbullying from tweets, we have used an already created benchmark dataset [16], from which we have randomly taken 9093 tweets. These data are stored as a CSV and each data file contains 2 columns- 'tweet' and 'class'. Tweet indicates the collection of all the tweets, both offensive and non-offensive, gathered for our experiment. And Class represents the labeling of each tweet. Tweets that are not offensive are labeled as '0' and the offensive ones are labeled

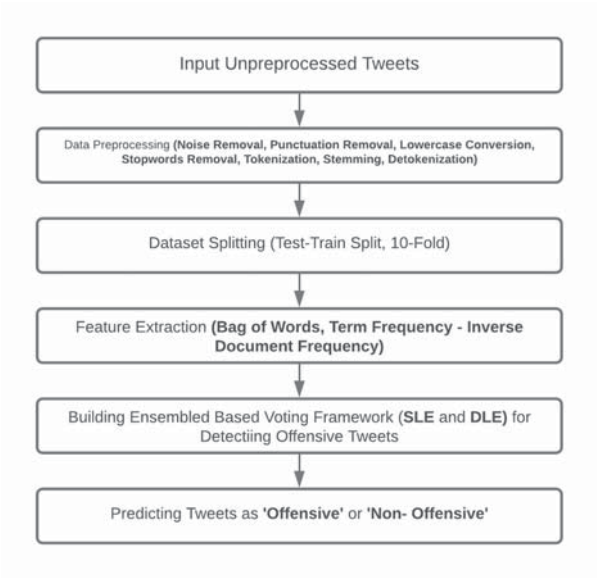

Fig. 1. Overall Cyberbullying Detection Process.

as '1'. 4164 tweets are labeled as non-offensive (45.8%) and the remaining 4929 tweets are tagged as offensive (54.2%).

B. Data Preprocessing

As the tweets collected from the dataset was in raw format, it needed to be refined to be ready for the implementation of our project. Several preprocessing techniques have been used to get the data in a clean format. The process included noise removal (tags, links, whitespace and numbers), punctuation removal followed by converting the whole document into lower case order for consistency. Then tokenization is performed after which insignificant words known as stopwords have been removed from the documents. These are the most common words used in a sentence that would have created noise if they were included as a feature in text classification. Consequently, to decrease the types of words in the document, we have converted the tokens to their root form which is known as stemming. We have used Snow Ball Stemmer for this purpose. Finally, in the last step in preprocessing, detokenization has been carried out which has resulted in clean and suitable data for advancing to the next stage. The overall preprocessing has been pictured in Figure 2.

C. Dataset Splitting

The cleaned label dataset has been split into a train (80%) and a test set (20%). Features have been fitted to the train set after completion of the splitting. Because our dataset was not fully balanced, we have validated our model with several cross-validation techniques: K-Fold, Stratified K-Fold, Shuffle Split and Stratified Shuffle Split. 10 folds have been used to apply these techniques.

D. Feature Extraction

Learning from a high volume of data in text classification is quite challenging because of the number of words, terms and phrases. Thus, the whole process becomes computationally very expensive. Additionally, irrelevant and repetitive features impede the accuracy and performance of any classification model. Hence, it is quite optimal to extract features to reduce the large volume of data size and avoid working with high dimensional data. We have studied in this research with two techniques to extract features: Bag of Words and TF-IDF. The Bag of Words (BoW) technique takes each tweet as input and calculates the number of occurrences of each word in that tweet, which then generates a numerical representation of the text known as vector features of fixed length. Thus, raw words are encoded into a count vector format as key-value pairs containing the frequency of each word in the text. In addition to that, We have also used the "Term Frequency-Inverse Document Frequency" (TF-IDF) technique for information retrieval in our project. TF-IDF is a statistical measure that assigns numerical weights to text data that is used in mining. It calculates the relative significance of a tern within the whole dataset by measuring the number of occurrences of that term in text data. The main characteristic of this technique is that Inverse Document Frequency counteracts with Term Frequency. In other words, it scales down the impact of the most common words occurring in the string by weighing up the rare ones. In this work, we have used different TF-IDF analyzers such as: Word, Character, Unigram and Bi-gram.

E. Classification Process

To detect cyberbullying from social media, various machine learning algorithms have been considered. Naive Bayes and SVM did not provide good result [4]. Distinct text features as discussed in the previous subsection along with specific mining algorithms have been outlined in [9] where the performance of ensemble techniques is missing. In this paper, we have traced the performance of MNB, LR, DT, LSVC, GBoost, AdB, Bagging on our dataset with all the features mentioned previously. To increase the accuracy with other performance metrics, we have applied an ensemble of these ML algorithms. SLE and DLE of our performed ML algorithms have been proposed. Figure 3 portrays our proposed framework.

For SLE, we have created a voting classifier for all the algorithms we have employed for our detection scheme. In the case of DLE, we have partitioned different classifiers into two groups and two voting classifiers have been created.
$VC1$: MNB, LR, DT, LSVC; and $VC2$: GBoost, AdB, Bagging;

The DLE has been constructed as follows:
$1st\ Level$: $VC1$ and $VC2$ has been formed
$2nd\ Level$: $VC3$ has been resulted from $VC1$ and $VC2$

In DLE architecture, voting classifiers have been created in two levels. In the first level, two groups ($VC1$ and $VC2$) have been formed in such a way that the generic ML models can provide a voting result and the ensemble-based models can also predict a class. The result of $VC1$ is then compared with $VC2$ in the second level to get the final classification outcome.

Fig. 2. Preprocessing Steps.

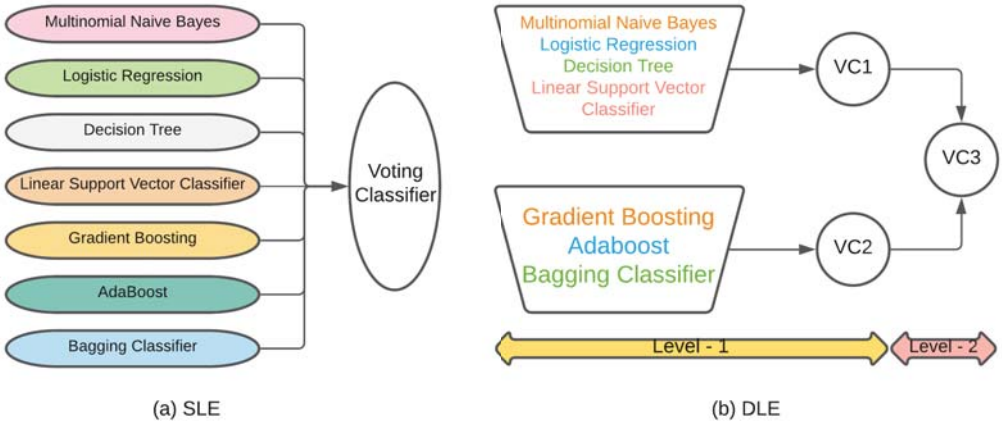

Fig. 3. Proposed SLE and DLE models.

In our proposed model, we have applied hard voting to create the ensemble frameworks. Because of the hard voting model, if the majority of the classifiers in a classifier group predict a tweet as 'offensive', the ensemble model will assess that tweet as 'offensive'. Thus, the prediction results provided by our proposed scheme can validate the result with more confidence than individually applied algorithms.

To predict whether the text is 'offensive' or 'non-offensive', after evoking 'BoW' and 'TF-IDF' from the cleaned data, the features have been deployed to our proposed SLE and DLE, and various performance measurement metrics have been measured to evaluate our model performance.

IV. Experimental Results

We have built level based ensemble models to detect cyberbullying from Twitter data. 'BoW', TF-IDF have been generated from noise-free data and fitted to our proposed model. Among various ML algorithms, four well-known classifiers (MNB, LR, DT, LinearSVC) and three ensemble methods (GBoost, AdB and Bagging) have been taken to build our model. To estimate the performance of our framework, we have measured various performance evaluation metrics (Accuracy, F1-Score and AUC). The table I has shown the performance distinction of our models from individual ML models.

We have achieved 94% accuracy for SLE when TF-IDF ('Word' and 'Unigram') is used to extract features which has surpassed the performance of other models. 70% accuracy has been attained when TF-IDF ('Bi-gram') is considered to retrieve features from tweets. We have also obtained salient performance for other features as well for our SLE model. Moreover, F1-Score and AUC results have been also satisfactory. We have achieved the lowest accuracy for TF-IDF ('Bi-gram') since N-gram models generally try to find the context of sentences based on bytes, words, syllables, or characters and there are less contextual texts in our dataset as it has been collected from Twitter. Moreover, we have tuned to 5000 as maximum features when taking N-gram as a feature

TABLE I
COMPARATIVE ANALYSIS OF VARIOUS CLASSIFIERS BASED ON PERFORMANCE EVALUATION.

Metric	Feature Extraction		MNB	LR	DT	LSVC	Gboost	Adb	Bagging	SLE	DLE
Accuracy	BOW		0.92	0.94	0.92	0.91	0.92	0.9	0.91	0.92	0.92
	TF-IDF	Word	0.87	0.92	0.91	0.92	0.92	0.89	0.92	**0.94**	0.92
		Character	0.89	0.91	0.88	0.92	0.92	0.9	0.92	**0.93**	**0.93**
		Unigram	0.88	0.93	0.92	0.93	0.92	0.91	0.93	**0.94**	0.92
		Bi-gram	0.71	0.71	0.7	0.73	0.6	0.57	0.73	0.7	**0.75**
F1-Score	BOW		0.92	0.93	0.92	0.91	0.91	0.9	0.91	0.92	0.92
	TF-IDF	Word	0.87	0.92	0.91	0.92	0.92	0.89	0.92	0.94	0.92
		Character	0.89	0.91	0.88	0.92	0.92	0.9	0.92	0.93	0.92
		Unigram	0.88	0.93	0.92	0.93	0.92	0.91	0.93	0.93	0.92
		Bi-gram	0.74	0.74	0.71	0.73	0.7	0.7	0.73	0.72	0.74
AuC	BOW		0.92	0.93	0.92	0.91	0.91	0.91	0.91	0.92	0.92
	TF-IDF	Word	0.88	0.92	0.91	0.92	0.92	0.9	0.92	0.94	0.92
		Character	0.89	0.91	0.88	0.92	0.92	0.9	0.92	0.93	0.93
		Unigram	0.89	0.93	0.92	0.93	0.93	0.91	0.93	0.93	0.92
		Bi-gram	0.78	0.78	0.75	0.76	0.76	0.76	0.76	0.75	0.77

TABLE II
PERFORMANCE EVALUATION OF SLE AND DLE MODEL BASED ON CROSS-VALIDATION TECHNIQUES.

Cross Validation Technique	Feature Extraction		SLE			DLE		
			Accuracy	F1-Score	AuC	Accuracy	F1-Score	AuC
K-fold	BOW		0.95	0.95	0.95	**0.96**	0.95	0.95
	TF-IDF	Word	0.95	0.95	0.95	**0.96**	0.94	0.96
		Character	0.95	0.95	0.95	**0.96**	0.94	0.95
		Unigram	**0.96**	0.96	0.96	**0.96**	0.95	0.96
		Bi-gram	0.68	0.81	0.76	0.71	0.60	0.76
Stratified K-fold	BOW		0.94	0.95	0.95	0.95	0.96	0.95
	TF-IDF	Word	0.95	0.95	0.95	0.95	0.96	0.97
		Character	0.94	0.95	0.95	**0.96**	0.96	0.96
		Unigram	0.95	0.95	0.95	**0.96**	0.97	0.96
		Bi-gram	**0.83**	0.86	0.86	0.79	0.79	0.79
Shuffle Split	BOW		0.94	0.95	0.94	0.94	0.94	0.94
	TF-IDF	Word	0.95	0.96	0.96	0.94	0.94	0.94
		Character	0.94	0.94	0.94	0.94	0.93	0.94
		Unigram	0.95	0.95	0.95	0.93	0.95	0.94
		Bi-gram	0.76	0.81	0.83	0.78	0.79	0.79
Stratified Shuffle Split	BOW		0.94	0.96	0.95	0.94	0.94	0.94
	TF-IDF	Word	0.94	0.96	0.95	0.94	0.94	0.94
		Character	0.95	0.95	0.95	0.94	0.93	0.94
		Unigram	**0.96**	0.95	0.95	0.93	0.94	0.94
		Bi-gram	0.78	0.82	0.84	**0.79**	0.78	0.79

value. We have also increased another level and divided the algorithms into two classifier groups for constructing the DLE model. The division has been made in such a way that all well-known ML classifiers are placed in one group and the ensemble methods are placed in another group. Though SLE has achieved higher performance than DLE, for TF-IDF('Bi-gram'), DLE has achieved 75% accuracy which is greater than SLE. Previously, for detecting hateful speech, the accuracy achieved was not up to the mark. In [4] [9] [15], the prediction accuracy varied from 52% to 90% which are overcome in our proposed model. Voting classifiers help to justify the class label as it depends on the majority voting scheme. For example, if five of the seven algorithms support a particular label, the overall classifier will take the majority supported result only. This technique is unique in this work for accomplishing the better results. DLE model has been applied for comparing the results with SLE. For two or three context-based word analysis is measured, DLE can outperform SLE too. Consequently, both architectures are useful in bullying detection. Ensemble based voting classifiers have been superior for bullying speech discernment.

Our dataset was partially balanced. To prevent our model from bias, we have also validated our model by applying different cross-validation techniques (K-Fold, Stratified K-Fold, Shuffle Split, Stratified Shuffle Split). 10-Fold cross-validation has been applied in this work. Stratified K-Fold and Stratified Shuffle Split have been considered to randomize the dataset into 10 folds so that class wise split can be maintained. The cross-validation result of our suggested model is depicted in table II.

We have reached 95% to 96% accuracy in the case of SLE and DLE respectively for all the features except 'Bi-gram' when the K-Fold cross-validation technique has been utilized. The highest 83% accuracy has been attained for TF-IDF ('Bi-gram') when Stratified K-Fold technique has been applied for SLE and Stratified Shuffle Split has yielded 79% accuracy for DLE. We have also accomplished good performance measurement metrics results for other features as

well. This cross-validation technique comparison was absent in our studied literature review.

In the DLE model, we have grouped the same category algorithms together which has allowed us to preserve objectivity. The results we have obtained are highly impressive. Voting based ensemble models are unique in this work and both have validated the prediction of tweet labels by achieving the highest 96% accuracy with an almost similar result for F1-Score and AUC. We have also calculated Precision and Recall for analysis and as F1-Score comes from these two performance metrics, that is why we have not mentioned them in the tables. After overall result analysis, we have finally come up with a decision that our machine learning based SLE and DLE can be a good approach to carry out the best result in the state of art. Taking the average performance for all the feature extraction techniques, a performance summary of all the aforementioned models based on three metrics namely, Accuracy, F1-Score and AuC, has been shown in a bar-chart below in the figure 4.

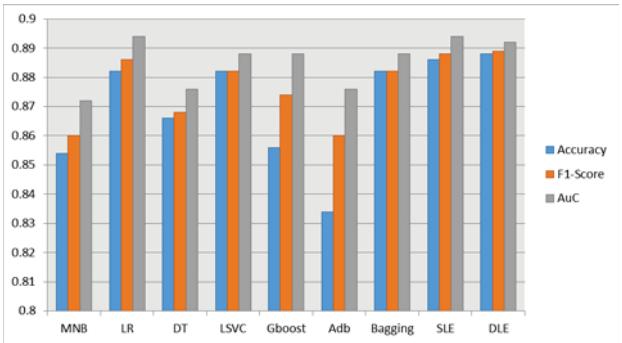

Fig. 4. Performance Summary of all the models.

In this figure, all the features have been taken to calculate the average accuracy, F1-Score and AuC value. Balanced performance metric values have been observed in our proposed ensemble-based voting architecture. As social bullying can make a huge impact on relationships among people and society, high accuracy in the detection process based on ML models is necessary. Our proposed SLE and DLE models can be reliable here for optimizing the classification process. Cross-validation techniques have also provided good performance in classifying social bullying. So these proposed models can be easily deployed in fact-checking websites for a better outcome in this context.

V. Conclusion & Future Works

Cyberbullying or Offensive messages and posts over social media are continuously affecting individuals particularly teenagers and society and often lead to a series of consequences even suicidal thoughts among the victims. In this research, we have built two ensemble based voting models to detect offensive or non-offensive texts. Our proposed model has outperformed all the independently applied ML algorithms and ensemble techniques. We have achieved the highest 96% accuracy for the twitter extracted dataset. In the future, we will try to collect multiple diversified datasets and some private datasets to measure the performance of our model. Finally, as indicated before, cyberbullying takes many forms such as harassment, flaming, denigration, impersonation, racism, sexism etc. So far, we have only categorized data into two groups. Therefore, it would be great to extend and examine if the proposed model can work for the multi-class classification problem too. In fact, our suggested models can be applied in other related text classification works for further meaningful analysis.

References

[1] Anna Schmidt & Michael Wiegand. (2017). A Survey on Hate Speech Detection using Natural Language Processing. 1-10. 10.18653/v1/W17-1101.

[2] Gengfeng Niu, Jing he, Shanyan Lin, Xiaojun Sun and Claudio Longobardi. 2020. "Cyberbullying Victimization and Adolescent Depression: The Mediating Role of Psychological Security and the Moderating Role of Growth Mindset" Int. J. Environ. Res. Public Health 17, no. 12: 4368.

[3] Abigail Geiger, "How and why we studied teens and cyberbullying," Pew Research Center, 2018. [Online]. Available: http://www.pewresearch.org/fact-tank/2018/09/27/qa-how-and-why-westudied-teens-and-cyberbullying/.

[4] R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893.

[5] Shan Suthaharan, "Machine learning models and algorithms for big data classification." Integr. Ser. Inf. Syst 36 (2016): 1-12.

[6] Mucahid Mustafa Saritas and Ali Yasar. "Performance analysis of ANN and Naive Bayes classification algorithm for data classification." International Journal of Intelligent Systems and Applications in Engineering 7.2 (2019): 88-91.

[7] Nabi Rezvani and Alireza Tabebordbar. "Linking Textual and Contextual Features for Intelligent Cyberbullying Detection in Social Media."

[8] Bandeh Ali Talpur and Declan O'Sullivan. "Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyberbullying in Twitter." Informatics. Vol. 7. No. 4. Multidisciplinary Digital Publishing Institute, 2020.

[9] Amgad Muneer and Suliman Mohamed Fati. "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter." Future Internet 12.11 (2020): 187.

[10] Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia. "Improving cyberbullying detection using Twitter users' psychological features and machine learning." Computers & Security 90 (2020): 101710.

[11] Mahen Herath, Thushari Atapattu, Hoang Dung, Christoph Treude, and Katrina Falkner. 2020. "AdelaideCyC at SemEval-2020 Task 12: Ensemble of Classifiers for Offensive Language Detection in Social Media. In Proceedings of SemEval".

[12] Nijia Lu, Guohua Wu, Zhen Zhang, Yitao Zheng, Yizhi Ren, and Kim-Kwang Raymond Choo. "Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts." Concurrency and Computation: Practice and Experience (2020): e5627.

[13] Rashi Shah, Srushti Aparajit, Riddhi Chopdekar, and Rupali Patil. "Machine Learning based Approach for Detection of Cyberbullying Tweets." International Journal of Computer Applications 975: 8887.

[14] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean. "A comparison of classification algorithms for hate speech detection." In IOP Conference Series: Materials Science and Engineering, vol. 830, no. 3, p. 032006. IOP Publishing, 2020.

[15] I. Alanazi, & J. Alves-Foss (2020). Cyber Bullying and Machine Learning: A Survey. International Journal of Computer Science and Information Security (IJCSIS), 18(10).

[16] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1. 2017.