Cyberbullying Detection on Social Networks Using Machine Learning Approaches

Md Manowarul Islam
Dept. of CSE
Jagannth University
Dhaka, Bangladesh
manowar@cse.jnu.ac.bd

Md Ashraf Uddin
School of Science, IT and Physical Sciences
Federation University Australia
Ballarat, Australia
ma.uddin@federation.edu.au

Linta Islam
Dept. of CSE
Jagannth University
Dhaka, Bangladesh
linta@cse.jnu.ac.bd

Arnisha Akter
Dept. of Computer Science and Engineering
Jagnnath University
Dhaka, Bangladesh
arnisha@cse.jnu.ac.bd

Selina Sharmin
Dept. of Computer Science and Engineering
Jagnnath University
Dhaka, Bangladesh
selina@cse.jnu.ac.bd

Uzzal Kumar Acharjee
Dept. of CSE
Jagnnath University
Dhaka, Bangladesh
ukacharjee@gmail.com

2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 978-1-6654-1974-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSDE50874.2020.9411601

***Abstract*—The use of social media has grown exponentially over time with the growth of the Internet and has become the most influential networking platform in the 21st century. However, the enhancement of social connectivity often creates negative impacts on society that contribute to a couple of bad phenomena such as online abuse, harassment cyberbullying, cybercrime and online trolling. Cyberbullying frequently leads to serious mental and physical distress, particularly for women and children, and even sometimes force them to attempt suicide. Online harassment attracts attention due to its strong negative social impact. Many incidents have recently occurred worldwide due to online harassment, such as sharing private chats, rumours, and sexual remarks. Therefore, the identification of bullying text or message on social media has gained a growing amount of attention among researchers. The purpose of this research is to design and develop an effective technique to detect online abusive and bullying messages by merging natural language processing and machine learning. Two distinct freatures, namely Bag-of - Words (BoW) and term frequency-inverse text frequency (TF-IDF), are used to analyse the accuracy level of four distinct machine learning algorithms.**

***Index Terms*—Cyberbullying, Machine learning, Natural language processing, Social media.**

I. Introduction

Social media is a platform that allows people to post anything like photos, videos, documents extensively and interact with society [1]. People connect with social media using their computers or smartphones. The most popular social media includes Facebook[1], Twitter[2], Instagram[3], TikTok[4] and so on. Nowadays, social media is involved in different sectors like education [2], business [3], and also for the noble cause [4]. Social media is also enhancing the world's economy through creating many new job opportunities [5].

[1]https://www.facebook.com/
[2]https://twitter.com/
[3]https://www.instagram.com/
[4]https://www.tiktok.com/

Although social media has a lot of benefits, it also has some drawbacks.Using this media, malevolent users conduct unethical and fraudulent acts to hurt others feelings and damage their reputation. Recently, cyberbullying has been one of the major social media issues. Cyberbullying or cyber-harassment refers to an electronic method of bullying or harassment. Cyberbullying and cyber-harassment are also known as online bullying. As the digital realm has grown and technology has progressed, cyberbullying has become relatively common, particularly amongst adolescents.

Approximately 50% of the teenagers in America experience cyberbullying [6]. This bullying has a physical and mental impact on the victim [7]. The victims choose self-destructive acts like suicide because the trauma of cyberbullying which is hard to be endured [8]. Thus, the identification and prevention of cyberbullying is important to protect teenagers.

In this context, we suggest a cyberbullying detection model based on machine learning that can detect whether a text relates to cyberbullying or not. We have investigated several machine learning algorithms, including Naive Bayes, Vector Machines for Support, Decision Tree, and Random Forest in the proposed cyberbullying detection model. We conduct experiments with two datasets from twitter and Facebook's comments and posts. For performance analysis, we use two different feature vectors BoW and TF-IDF. The results indicate that TF-IDF feature provides better accuracy than BoW where SVM provides better performance than any other machine learning algorithms used in this paper.

The remainder of the paper is organized as follows. Section II overviews the related works. Section III presents the details of the proposed machine learning-based model. Section IV shows the experiment results. Section V concludes the paper and highlights some future work.

II. Related Works

There are several works on machine learning-based cyberbullying detection. A supervised machine learning algorithm was proposed using a bag-of-words approach to detect the sentiment and contextual features of a sentence [9]. This algorithm shows barely 61.9% of accuracy. Massachusetts Institute of Technology conducted a project called Ruminati [10] employing support vector machine to detect cyberbullying of youtube comments. The researcher combined detection with common sense reasoning by adding social parameters. The result of this project was improved to 66.7% accuracy for applying probabilistic modelling. Reynolds et al. [11] proposed a language-based cyberbullying detection method which shows 78.5% of accuracy. The authors used the decision tree and instance-based trainer to achieve this accuracy. To improve cyberbullying detection, the author of the paper [12] has used personalities, emotion and sentiment as the feature.

Several deep learning-based models were also introduced to detect the cyberbullying. Deep Neural Network-based model is applied for cyberbullying detection by using real-world data [13]. The authors first analyze cyberbullying systematically then used transfer learning to do the detection task. Badjatiya et al. [14] has presented a method using deep neural network architectures for detecting hate speech. A convolutional neural network-based model has been proposed to detect cyberbullying [15]. The authors employed word embedding where similar words have similar embedding. In a multi-modal context, Cheng et al. [16] research the novel issue of cyberbullying identification by collaboratively exploiting social media data. This challenge, however, is difficult due to the complex combination of both cross-modal associations among multiple methods and structural correlations between various social media sessions, and the complex attribute information of different modalities. They propose XBully, a novel cyberbullying identification system to overcome these challenges, which first reformulates multi-modal social media data as a heterogeneous network and then tries to learn node embedding representations on it.

Many literatures on cyberbullying have concentrated on text analysis over the past few decades. Cyberbullying, however, is becoming multi-objective, multi-channel, and multi-form. The variety of bullying data on social platforms can not be met by conventional text analytical techniques.

Wang et al. [17] suggested a multi-modal identification system that integrates multi-modal information such as image, video, comments, time on social media to cope with the latest type of cyberbullying. In particular, they not only extract textual characteristics, but also apply hierarchical attention networks to capture the social network session function and encode various media information including video, image. The authors model the multi-modal cyberbullying detection system to address the latest type of cyberbullying on the basis of these characteristics.

Using Neural Networks to facilitate the identification of online bullying has become common in recent years. These Neural Networks are also based solely on or in conjunction with other layer types utilising Long-Short-Term-Memory layers. Buan et al. [18] introduced a new model for the Neural Network that can be applied in textual media to identify evidence of cyberbullying. The concept is made on existing architectures that merge the strength of Long-Short-Term-Memory layers with Convolutionary layers. In addition, their architecture features the use of stacked core layers, which demonstrates that their study enhances the Neural Network's efficiency. A new type of activation method is also included in the design, that is called "Support Vector Machine like activation" By using L2 weight regularisation and a linear activation function in the activation layer along with using a Hinge loss function, the "Support Vector Machine like activation" is accomplished.

By creating a machine learning system with three distinctive features, Raisi et al. [19] resolve the computational problems related to harassment identification in social networks. (1) In the form of specialist-provided key phrases that are predictive of bullying or non-bullying, minimal supervision is applied. (2) With an aggregate of two learners who co-train one another, this identifies bullying; one learner investigates the language content in the text and the other learner recognizes the social structure. (3) By training nonlinear deep models, this integrates decentralized word and graph-node representations. By optimising an objective function that combines a co-training loss with a weak-supervision loss, the model is trained.

Cyberbullying has recently been identified by users of online social networks as a significant national health problem and the creation of an effective detection model has considerable scientific merit. Al et al. [20] have introduced a collection of specific Twitter-derived features including behaviour, user, and tweet content. They have built a supervised machine learning solution for the detection of cyberbullying on Twitter based network. An assessment shows that, based on their proposed features, their established detection system obtained outcomes with a region under the receiver-operating characteristic curve of 0.943 and an f-measure of 0.936.

For those affected, cyberbullying can escalate to deep psychological and mental issues. There is also an immediate need to formulate automated approaches for the identification and prevention of cyberbullying. Although recent cyberbullying detection efforts have established advanced methods of text processing for cyberbullying detection, there are still few attempts to use visual data processing to detect cyberbullying automatically. Singh et al. [21] reported that image elements support feature vectors in cyberbullying detection based on early analysis of a public, labelled cyberbullying dataset, and can actually boost predictive performance. When cyberbullying is becoming more and more common in social networks, it becomes of extreme significance to immediately identify and reactively respond upon this. The work in [22] researched how Fuzzy Fingerprints, a recent technique with documented effectiveness in comparable tasks, works while identifying textual cyberbullying in social networks.

III. Bullying Detection Model

In this section, we describe the cyberbullying detection framework which consists of two major parts as shown in 1. The first part is called NLP (Natural Language Processing) and the second part is named as ML (Machine learning). In the first phase, datasets containing bullying texts, messages or post are collected and prepared for the machine learning algorithms using natural language processing. The processed datasets are then used to train the machine learning algorithms for detecting any harassing or bullying message on social media including Facebook and Twitter.

A. Methodology

- **Natural Language processing:** The real world posts or text contain various unnecessary characters or text. For example, numbers or punctuation are irrelevant to bullying detection. Before applying the machine learning algorithms to the comments, we need to clean and prepared them for the detection phase. In this phase, various processing task including removal of all irrelevant characters like stop-words, punctuation and numbers, tokenizations, stemming etc. After the preprocessing, we prepare the two important features of the texts as follows:
  1) Bag-of-Word: The machine learning algorithms cannot work directly with the raw text. So before applying the algorithms we must convert them to vectors or numbers. So, the processed data is converted to Bag-of-Words (BoW) for the next phase.
  2) TF-IDF: This is another features that we consider for our model. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that can evaluate how relevant a word is to a document in a collection of documents. In bag of words, every word is given equal importance while in TF-IDF the words that occur more frequently should be given more importance as they are more useful for classification.
- **Machine Learning:** This module involves in applying various machine learning approaches like Decision Tree (DT), Random Forest, Support Vector Machine, Naive Bayes to detect the bullying message and text. The classifier with the highest accuracy is discovered for a particular public cyberbullying dataset. Next section, some common machine learning algorithms are discussed to detect cyberbullying from social media texts.

B. Machine Learning Algorithms

In this section, we discussed the basic mechanisms of several machine learning algorithms. We presented Decision Tree, Naive Bayes, Random Forest and Support Vector Machine in each subsection.

1) Decision Tree: The decision tree classifier can be used in both classification and regression [23]. It can help represent the decision as well as make a decision. The decision tree is a tree-like structure where each internal node represents a condition, and each leaf node represents a decision. A classification tree returns the class where the target falls. A regression tree yields the predicted value for an addressed input.

2) Naive Bayes: Naive Bayes is an efficient machine learning algorithm based on Bayes theorem [24]. The algorithm predicts depending on the probability of an object. The binary and multi-class classification problems can be quickly solved using this technique. Based on Bayes' Theorem it finds the probability of an event occurring given the probability of another event that has already occurred as follows:

$$p(y|X) = \frac{p(X|y) \times p(y)}{X} \tag{1}$$

Here, where, the class variable is denoted by y and X is a dependent feature vector of length n as $X = x_1, x_2, x_3 ... x_n$.

3) Random Forest: Random Forest classifier is consists of multiple decision tree classifiers [25]. Each tree gives a class prediction individually. The maximum number of the predicted class is our final result. This classifier is a supervised learning model which provides accurate result because several decision trees are merged to make the outcome. Instead of relying on one decision tree, the random forest takes the prediction from each generated tree and based on the majority votes of predictions, and it decides the final output. For example, if we have two classes namely A and B and the most of the decision tree predict the class label B of any instance, then RF will decides the class label B as follows:

$$f(x) = \text{majority vote of all tree as B} \tag{2}$$

4) Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm which can be applied in both classification and regression alike a decision tree. It can distinguish the classes uniquely in n-dimensional space [26]. Thus, SVM produces a more accurate result than other algorithms in less time. In practice, SVM constructs set a of hyperplanes in a infinite-dimensional space and SVM is implemented with kernel which transforms an input data space into the required form. For example, Linear Kernel uses the normal dot product of any two instances as follows:

$$K(x, x_i) = \text{sum}(x * x_i) \tag{3}$$

IV. Experiment and Results

We have used the four machine learning algorithms namely, Decision Tree (DT), Naive Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF) to classify comments as bullying or non-bullying. In this section, we first describe the datasets for the experiment and then discuss about the results.

A. Datasets

We have collected Facebook comments from different posts (Dataset-1) and the twitter comments dataset from kaggle.com [27] for (Dataset-2). The texts or comments were classified into two types as follows:

- **Non-bullying Text:** This type of comments or posts are non-bullying or positive comments. For example, the

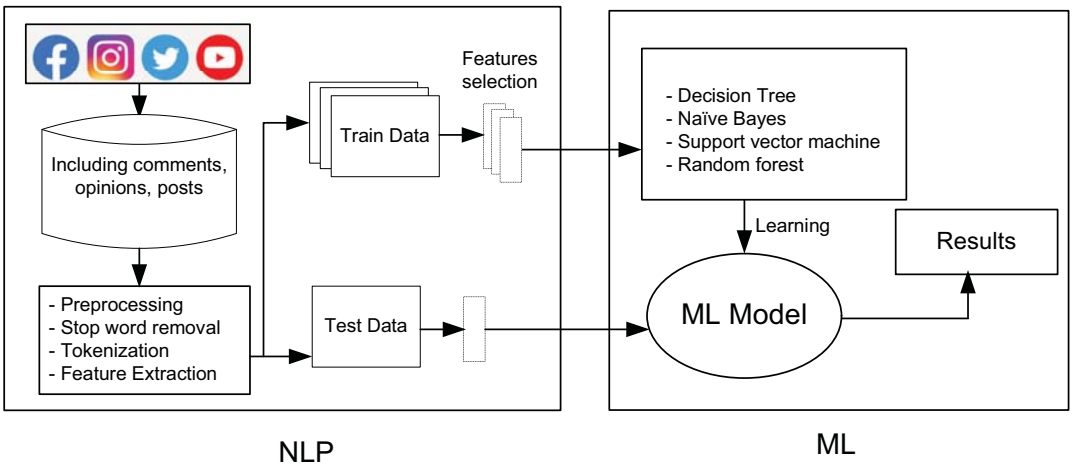

Fig. 1. Proposed framework for bully detection

comment like "This photo is very beautiful" is positive and non-bullying comments.

- **Bullying Text:** This type belongs to bully type comments or harassment's. For example, "go away bitch" is a bullying text or comment and we consider as negative comment.

The bullying detection algorithms are implemented using python machine learning packages. The performances are analyzed with respect to the following metrics.

- The classification results are listed in the confusion matrix [28] shown in Table I, also called the contingency table. The True Positive upper left corner is the number of individuals that were listed as true positive, while those were true. The False-positive lower right cell reflects the number of samples that, though false, were labelled as false negative. False-negative shows the number of individuals, while these were false, being counted as true. False-positive reflects, as these were true, the number of individuals that were listed as true.

$$\text{Accuracy} = \frac{\sum \text{True Positive} + \sum \text{True Negative}}{\sum \text{Total Samples}} \quad (4)$$

TABLE I
THE CONFUSION MATRIX

	Condition Positive	Condition Negative
Predicted Condition Positive	True Positive	False Negative
Predicted Condition Negative	False Positive	True Negative

$$\text{Precision} = \frac{\sum \text{True Positive}}{\sum \text{Predicted Condition Positive}} \quad (5)$$

$$\text{Recall} = \frac{\sum \text{True Positive}}{\sum \text{Condition Positive}} \quad (6)$$

- Receiver The Operating Characteristic Curve (or ROC Curve) [29] is a plot of the true positive rate for the various potential diagnostic test cutpoints against the false-positive rate. The trade-off between sensitivity and specificity is exposed by ROC (a reduction in specificity would follow any increase in sensitivity). The more the curve follows the left border and the more closely the curve follows the ROC space 's top border, the more precise the test would be.

What follows we are describing the results of the proposal.

B. Results for Dataset-1

This dataset is built from the user comments on different Facebook posts. We compare the various parameters of the machine learning algorithms based on the two important features vectors BoW and TF-IDF. Figure 2 and 3 show the precision and accuracy results and from the graph it is clear that SVM outperforms the other algorithm. The results also indicate that, TF-IDF provides better accuracy than BOW feature. This is because rather than taking almost all word into vectors, TF-IDF takes the most frequent words and maintain better performances.

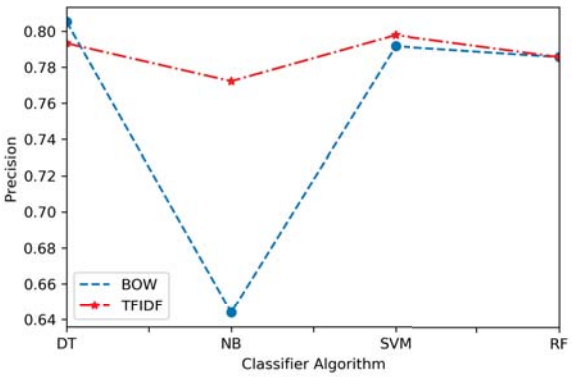

Fig. 2. Precision for Dataset-1

Figure 4 and 5 represents the Receiver Operating Characteristics (ROC) curve for both features. For BoW and TF-IDF

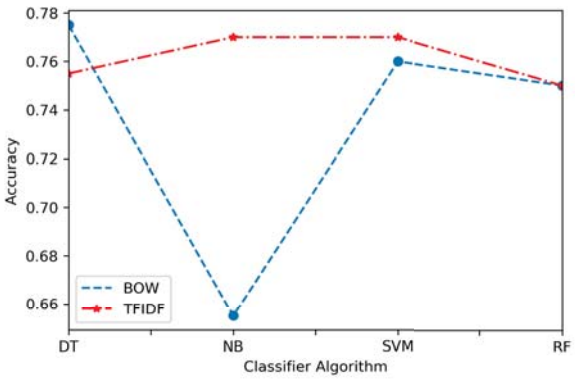

Fig. 3. Accuracy for Dataset-1

it is clear that SVM provide higher performance than the other classifier algorithms.

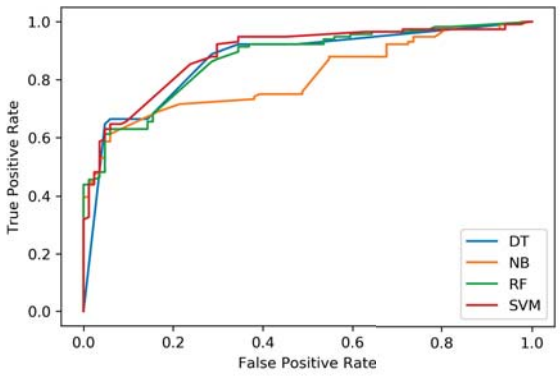

Fig. 4. ROC curve for BoW

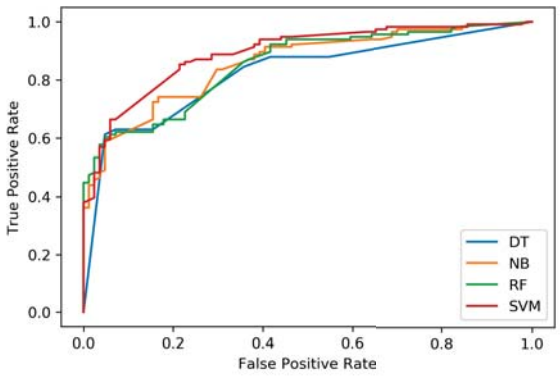

Fig. 5. ROC curve for TF-IDF

C. Results for Dataset-2

The precision and accuracy graph of various machine learning algorithms are shown in Figure 6 and 7. We observed the similar results and found that TF-IDF provide better accuracy performances than BoW. Among the machine learning algorithms SVM outperforms the others.

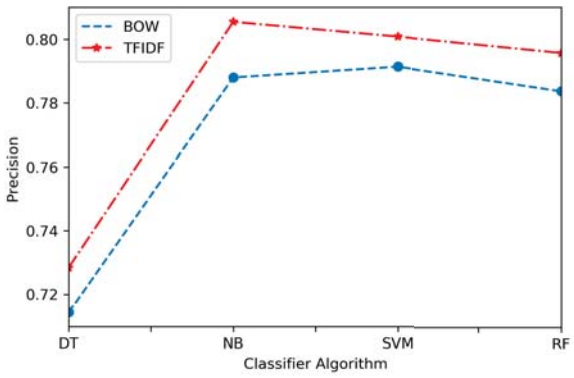

Fig. 6. Precision for Dataset-2

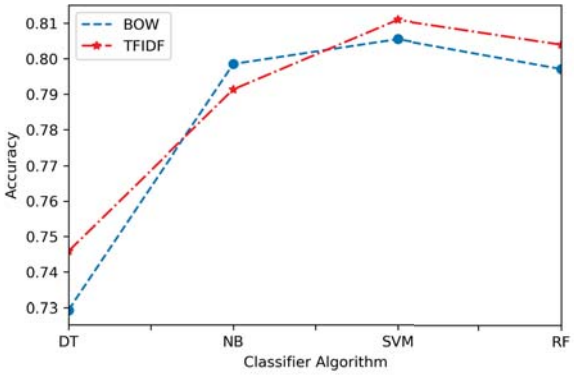

Fig. 7. Accuracy for Dataset-2

Figure 8 and 9 shows the ROC curve for both BoW and TF-IDF and from the graphs it is clear that SVM exhibits better performance accuracy compare to the other classifier algorithms.

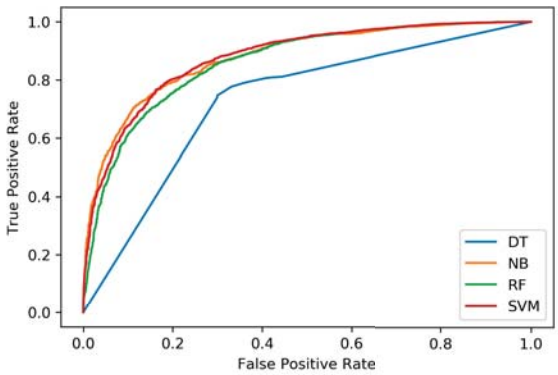

Fig. 8. ROC curve for BoW

V. Conclusion

In particular, cyberbullying has become more common and has begun to raise significant social issues with the rising prevalence of social media sites and increased social media use by teenagers. There needs to design automatic cyberbullying

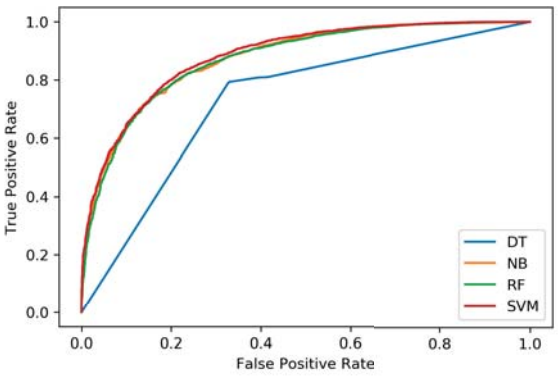

Fig. 9. ROC curve for TF-IDF

detection method to avoid bad consequences of cyber harassment. Considering the significance of cyberbullying detection, in this study, we investigated the automated identification of posts on social media related to cyberbullying by considering two features BoW and TF-IDF. Four machine learning algorithms are used to identify bullying text and SVM for both BoW and TF-IDF. In future we are planning to design a framework for automatic detection and classification of cyberbullying from Bengali texts using deep learning algorithms.

ACKNOWLEDGMENT

This work is supported by Jagannath University Research grant.

REFERENCES

[1] C. Fuchs, *Social media: A critical introduction*. Sage, 2017.
[2] N. Selwyn, "Social media in higher education," *The Europa world of learning*, vol. 1, no. 3, pp. 1–10, 2012.
[3] H. Karjaluoto, P. Ulkuniemi, H. Keinänen, and O. Kuivalainen, "Antecedents of social media b2b use in industrial marketing context: customers' view," *Journal of Business & Industrial Marketing*, 2015.
[4] W. Akram and R. Kumar, "A study on positive and negative effects of social media on society," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 10, pp. 351–354, 2017.
[5] D. Tapscott *et al.*, *The digital economy*. McGraw-Hill Education,, 2015.
[6] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. Desmet, and I. De Bourdeaudhuij, "Cyberbullying on social network sites. an experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully," *Computers in Human Behavior*, vol. 31, pp. 259–271, 2014.
[7] D. L. Hoff and S. N. Mitchell, "Cyberbullying: Causes, effects, and remedies," *Journal of Educational Administration*, 2009.
[8] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of suicide research*, vol. 14, no. 3, pp. 206–221, 2010.
[9] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," *Proceedings of the Content Analysis in the WEB*, vol. 2, pp. 1–7, 2009.
[10] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *In Proceedings of the Social Mobile Web*. Citeseer, 2011.
[11] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine learning and applications and workshops*, vol. 2. IEEE, 2011, pp. 241–244.
[12] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," *Computers & Security*, vol. 90, p. 101710, 2020.
[13] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*. Springer, 2018, pp. 141–153.
[14] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 759–760.
[15] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, 2018.
[16] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 339–347.
[17] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.
[18] T. A. Buan and R. Ramachandra, "Automated cyberbullying detection in social media using an svm activated stacked convolution lstm network," in *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, 2020, pp. 170–174.
[19] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 479–486.
[20] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.
[21] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2090–2099.
[22] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2018, pp. 1–7.
[23] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
[24] I. Rish *et al.*, "An empirical study of the naive bayes classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41–46, 2001.
[25] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.
[26] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.
[27] "Datasets," https://www.kaggle.com/datasets, accessed: June 2020.
[28] M. A. Uddin, A. Stranieri, I. Gondal, and V. Balasubramanian, "Rapid health data repository allocation using predictive machine learning," *Health Informatics Journal*, p. 1460458220957486, 2020.
[29] M. Ashraf Uddin, A. Stranieri, I. Gondal, and V. Balasubramanian, "Dynamically recommending repositories for health data: a machine learning model," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2020, pp. 1–10.