Abhorrent Tweet Text Detection Using Ensemble Of Classifiers Methods

Tanu Chouksey[1], Prof. Priyanka Saxena[2]
[2] Prof
[1, 2] Takshila Institute of Engineering & Technology, Jabalpur, M.P.

Abstract- *Nowadays Abhorrent Tweet on social media has become a major problem. Abhorrent tweet may cause many serious and negative mental, emotional and physical impacts on a person's life. However, abhorrence leaves a record that can demonstrate value and give proof to help stop digital abuse. The early detection of abhorrent tweet on social media becomes crucial to moving the effect on the social media user. Numerous studies are being done to automatically identify cyberbullying content in this trend. The absence of linguistic resources, especially for recently developed languages, is the main issue and gap in Abhorrent/Cyberbullying detection measures.*

Using Machine Learning with Natural Language Processing (NLP) techniques to automatically detect cyberbullying is the best way to stop it. Current research develops an efficient framework to detect Cyberbullying, using NLP tools with Machine Learning and Ensemble models. Using different preprocessing techniques, the proposed study is validated on an english-abusive-comment-detector. Five machine learning models Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) and their different combination (Ensembles) are evaluated on the different dataset. From experiments, the current study finds that the Ensemble model outperformed and achieved promising results from individual models. In last, an ensemble of these outperformed models is formed and achieved higher test accuracy.

Keywords- Abhorrent Tweet, Social Media, NLP, Machine Learning, Ensemble Model, Accuracy.

I. INTRODUCTION

The internet has become an important development tool for young people. It provides a great source of information and a tool for communication. In recent studies, children and young people categorized their Internet activities into three groups:

(a) Content-based activities, such as school work, play games, watch video clips, read the news, or download music;
(b) Contact/communication-based activities such as instant messaging, email, chatting or Skype; and
(c) Conduct peer participation activities such as blogging, post photos or file sharing websites [1].

Despite all the benefits, the Internet could be an environment for bullying. In their research, Haddon and Livingstone [2] showed that 17% of the children, who were interviewed between the age of 9 and 14 in the UK, were exposed to sexual content compared to 24% of children from the EU. The study also showed that the children experienced bad language in the form of insults or swearing, aggressive communication, or harassment. Moreover, social media platforms provide a fruitful environment for cyberbullying in the forms of threats, harassment, and exploiting potential victims [3]. The Pew research center reported in 2017 that 40% of social media users have experienced some form of cyberbullying [4]. Another study that included university students found that among 200 university students, 91% experienced cyberbullying, 55.5% of them on Instagram, and 38% on Facebook [5].

Cyberbullying experiences can have serious consequences for the victims, including depression, anxiety, low self-esteem, and self-harm, and may even lead in extreme cases to suicide [6]. Consequently, having tools for detecting and preventing cyberbullying is crucial for reducing the negative effects. Studying cyberbullying is rooted in Psychology, Education, Behavioural Science (BS), and Information Technology (IT). On the IT front, the automated detection of cyberbullying can help in the automated removal of the flagged content, post, or communication, in the automated blocking of the perpetrators, and in reaching out to help the victims.

Over the last decade, the body of literature on automated detection of cyberbullying has been growing, especially on the topic of detecting cyberbullying from social media networks like Twitter [7], Instagram [8], and YouTube [9]. This body of research has been working towards automated cyberbullying detection using either rule-based models [10], [11], conventional machine learning models [12],[13], or deep learning models [14]. The last decade

brought significant advances in the fields of Machine Learning (ML) and Natural Language Processing (NLP), which have been successfully applied in domains related to cyberbullying detection, such as rumor detection, sentiment analysis, and fake news detection. Consequently, it is extremely useful to review the available literature on automated cyberbullying detection, in light of these recent advances and proposed a model that will detect the hate tweets with improved accuracy and efficiency.

1.1 Cyberbullying

The lack of a globally accepted definition of cyberbullying is one of the main issues detected in the reviewed literature on automated cyberbullying detection. For example, although some of the reviewed works claim to detect cyberbullying in their title, they detect child grooming or detect the participants in the act, like the bullies, victims, and bystanders, rather than the actual incident of cyberbullying [15], [16]. There are some types of bullying that most of the papers used, as shown in Figure 1.1 below:

Figure 1.1: Cyberbullying Types.

However, despite that being close in meaning, as most of them describe cyberbullying as ``one form or another of insulting, spread using mobile or internet technology", the lack of a clear definition leads to difficulties in comparing and evaluating different works. For example, in [17], [18], cyberbullying is described as online aggression, bullying using new communication technologies, online harassment, or hate speech. This is problematic as each of these tasks is different, making it significantly difficult to replicate the studies and to compare the models' results and generalizability.

Some studies consider cyberbullying as a sub-type of cyber-aggression [19], while others consider cyberbullying as a different task from cyber-aggression [20]. Mladenovie *et al.* provided a detailed survey on the diversity of the definitions of cyberbullying, cyber-aggression, trolling, and cyber-grooming [21]. Another issue is that some studies do not differentiate between bullying and cyberbullying apart from the usage of electronic means. As a consequence, they require the following three characteristics of bullying to be evident in cyberbullying cases: harmful, repetitive, and with power imbalance between the bully and the victim. These characteristics sometimes are hard to satisfy in the online space. For example, someone may send a bullying message to someone during an online conversation only once, which does not satisfy repetition. However, some studies claim that the fact that an online post makes permanent harm satisfies the repetition requirement [22]. In addition, in the case of the Twitter platform, Tian and Xin argue that negative messages on Twitter tend to be retweeted more often, which also satisfies the repetition requirement [23].

1.2 Cyberbullying Types

According to the literature, there are 12 types of cyberbullying [24]:

1) Flamming: Starting a fight online.
2) Harassment: Sending insulting messages frequently.
3) Cyberstalking: Sending intimidating messages to the victim, which causes fear.
4) Masquerade: The bully pretends to be someone else.
5) Trolling: Posting controversial comments to upset other members on the online platform.
6) Denigration: Negative gossip about another person.
7) Outing: Posting personal information about someone in public forums.
8) Exclusion: When a social group deliberately excludes.
9) Cat_shing: Creating a fake profile using someone else's information.
10) Dissing: Posting information about someone to hurt them or defame them.
11) Trickery: Tricking someone to share their secrets or personal information.
12) Fraping: Using someone else's online account to post inappropriate content and tricking others into believing that the account owner posted them.

In the last few years, research on hate speech detection has been increasing [25], [26], [27]. In a survey paper on the automated detection of hate speech in text, Fortuna and Nunes studied the definition of hate speech in the literature in relation to four dimensions: physical violence encouragement, targets, attack language, and humorous hate speech. From these four dimensions, the authors proposed a new definition for hate speech, i.e. ``Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with

different linguistic styles, even in subtle forms or when humour is used".

In the NLP community, it is unclear what the difference in definition between hate speech and cyberbullying is. This lack of clarity can cause generalizability problems with the developed models, as each of the cyberbullying detection and hate speech detection tasks require different features. However, there are also some similarities between the two tasks. The main similarity is the abusive language, while the main difference is the target of the abusive language. In cyberbullying, the abusive language is targeted at specific individuals, while hate speech is targeted at groups of people who share specific characteristics [28]. Detection of cyberbullying requires intelligent systems because it is difficult to understand the complexities involved with text classification. Many machine learning models have been developed so far but deep learning models have not been exploited to the fullest in this domain. Moreover, current ML models work fine on a single social media platform (SMP) but fail when the same model is used on a different SMP. There are three ways by which cyberbullying detection can be done [29].

II. LITERATURE SURVEY

We broadly categories features used across the studies into 4 main groups, namely; content-, sentiment-, user and network-based features. We define content-based features as the extractable lexical items of a document such as keywords, profanity, pronouns and punctuations.

These are those features that are indicative of emotive content; they are generally keywords, phrases and symbols (e.g. emoticons) that can be used to determine the sentiments expressed in a document. User based features are those characteristics of a user's profile that can be used to make a judgments on the role played by the user in an electronic exchange and include age, gender and sexual orientation and finally, network-based features are usage metrics that can be extracted from the online social network and include items such as number of friends, number of followers, frequency of posting, etc.

The work of [30] examined several ML classifiers using various feature extraction and selection techniques on a dataset of YouTube comments in Arabic to detect offensive language in online communications. They applied a variety of feature transformation techniques applied, including logistic regression with L1 regularisation (LR-L1), feature ranking with recursive feature elimination (RFE), ExtraTreesClassifier, tree-based ensemble methods, and singular-value decomposition (SVD). They trained five classical ML classifiers using the extracted features: SVM, Naive Base (NB), Decision Tree (DT), and Logistic Regression (LR). They achieved the best results from the SVM classifier with combined features selected by LR-L1 and RFE. The accuracy, precision, recall, and F1-score rates were 84%, 89%, 76%, and 81%, respectively.

The work of [31] produced the first abusive dataset in Turkish called Abusive Turkish Comments (ATC) to detect offensive comments from the Instagram platform. It is composed of 10,528 abusive and 19,826 not-abusive comments. They built ML classifiers to detect abusive messages, such as NB, SVM, and XGBoost. They also used Convolutional Neural Network (CNN) in their detection system. The CNN model achieved the best micro-averaged F1-score rate of 97.4%.

Reducing the dimension of the features space becomes a vitally important step in the classification process since not all features are relevant for the classification task. In addition, the large number of features compared with the number of instances may lead to over-fitting [32]. Evolutionary algorithms, especially GA, are considered a perfect solution to explore the feature space. It can generate numerous features subsets during reproduction operations to get the best subset that comprises the most relevant features. For instance, [32] proposed an approach for Arabic opinions analysis, which combines SVM with a random subspace (RSS) algorithm, and applies GA to enhance the system. RSS is used to automatically generate different subsets of features vectors with limited size and replace the decision tree base classifier of RSS with SVM. The GA was applied to enhance the proposed methodology by avoiding the random choice adopted by RSS by generating features based on correlation criteria to avoid choosing incoherent features subsets. They trained the sentiment classifier on a corpus consisting of 1,000,000 Arabic reviews collected from online websites of Arabic Algerian newspapers. The enhancement made through using GA increased the accuracy rate of the proposed sentiment analysis system from 75.90% to 85.99%.

We group features such as cyberbullying keywords, profanity, pronouns, n-grams, Bags-of-words (BoW), Term Frequency Inverse Document Frequency (TFIDF), document length and spelling as content-based features. Content-based features are overwhelmingly used across as many authors utilising content-based features. As cyberbullying messages are often abusive and insulting in nature, it is not surprising that profanity was found to be the most used content based feature.

Studies such as Dinakar et al. [33], Perez et al. (2012) [34], Kontostathis et al. (2013) [35], Nahar et al. (2013) [36] and Bretschneider et al. (2014) [37], created profanity lexicons using word lists compiled by the researchers or sourced from external libraries such as noswearing.com3 and urbandictionary.com. By equating the presence of profanity to cyberbullying, the use of profanity lexicons alone fails to consider other key aspects of cyberbullying such as repetitiveness and the presence of a power differential.

Rafiq et al. (2015) [38] similarly cautioned against the use of profanity as the only feature for cyberbullying detection and argued that not all cyber aggression constitutes bullying. Studies such as Nahar et al. (2013) [36], Dadvar et al. (2014) [39], Bretschneider et al. (2014) [37] and Nahar et al. (2013) [36] incorporated other features such as pronouns alongside profanity to further detect instances where profane words are used in close proximity to a pronoun as such personalized abusive content are potentially more indicative of cyberbullying than the abusive terms on their own. For example, the phase "the f**king train was delayed again" is definitely not cyberbullying though it contained profanity but "you f**king idiot" could be. While this is an improvement, the pronoun + profanity feature still suffers the same shortcomings as using profane terms only. Dinakar et al. (2011) [40], often cited for the performance gain achieved by their label-specific binary classifiers over multi-class classifiers, achieved this improved performance by using domain-specific content features learned from training classifiers on a set of messages clustered on sensitive topics such as race, culture, sexuality and intelligence to then detect bullying messages within each cluster.

While Yin et al. (2009)[41] did not find n-grams very effective in their experiments, its use as detection feature is still relatively popular amongst studies including Dinakar et al. (2011), Xu et al. (2012) [42], Sood and Churchill (2012) [43], and Munezero et al. (2014) [44]. As TFIDF provides a measure of a word's importance to a document within a collection of documents, it can sometimes provide better results than using n-grams in isolation (Yin et al., 2009) [41] and it is therefore often used alongside n-gram and other features to improve detection performance as can be seen in the works of Yin et al. (2009), Dinakar et al. (2011)[33], Dadvar and De Jong (2012), , Sood and Churchill [43], and Nahar et al. (2013).

Sentiment-based Features Sentiment or emotion analysis has been used in areas such as detecting sentiments in informal product reviews on social media and analyzing market trends in financial forecasting. Within the field of cyberbullying detection, sentiment analysis is often combined with features like TFIDF and pronoun usage to improve the performance of the detection system. This is due to the fact that, while strong emotions can often be an indicator of bullying, they are rarely sufficient on their own to accurately identify a bullying episode. For example, a sarcastic sentence such as "I'm in love with your big nose" that scores high on positive emotions may also constitute bullying and would require additional methods to identify the phrase "big nose" as an instance of a potentially negative remark about an individual's physical appearance. If, however, within the same sentence "nose" is replaced by "eyes", this may very well be a declaration of affection or genuine admiration.

Further analysis of the tweets revealed, however, that fear is often expressed jokingly (e.g., "oooh I'm so scared") thus providing further evidence that a detection system based on sentiments only cannot always accurately distinguish between genuine emotions and those sarcastically expressed. This is in agreement with Dinakar et al. (2011)'s [40] discovery that bullying involving deliberate abuse and profanity were much easier to detect that those containing sarcasm and euphemism.

Munezero et al. [51] expanded the method by introducing two emotion based features directed at exploiting the emotional context of a post. The first emotion feature used ontology of emotions and emotive words based on WordNetAffect to determine the emotions expressed within text. The inclusion of these emotion based features improved the detection process in the majority of the experiments.

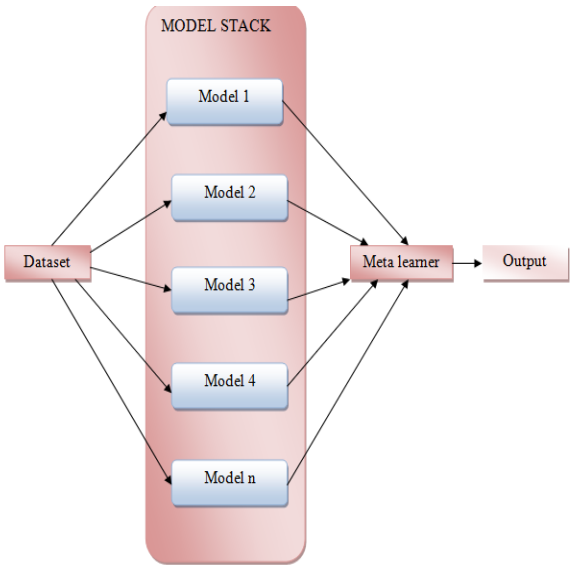
User-based Features Alongside content-based and emotion-based features, researchers have explored incorporating user-related features into cyberbullying detection systems. These include features like age, gender, sexual orientation and race. Dadvar and De Jong (2012) and Dadvar et al. used the TFIDF of profane words and pronouns as features in a gender-specific corpus of MySpace posts to train an SVM classifier. They found cyberbullying detection was significantly improved by the inclusion of gender-specific features when compared against results obtained using the same classifier trained on a no segregated dataset. While the improvements demonstrated by the study provide encouragement for the incorporation of gender features in online bullying detection, it should be noted that gender (and any other user supplied) information on social media can be easily falsified, therefore, any method that makes use of user supplied information will greatly benefit from means of validating such data – for example, a forensic linguistic module could be used to assign a "truth score" to age and gender information supplied by a user.

Network-based Features With the huge popularity of social media, its status as the predominant source of data for cyberbullying detection research, it is not surprising that network data such as number of friends, uploads, likes and so on are increasingly being used as features in detection systems. Nalini Priya and Asswini (2015) used the ego network to compute temporal changes in the relationships between users and use the detected changes within the detection process. Dadvar et al. used membership duration, number of uploads, subscriptions and comments posted as features in and activity history in alongside user-based and content-based features to achieve improved detection compared to experiments without network-based features.

While this method is heavily reliant on the human knowledge contained within its knowledge base, it certainly offers an avenue to improve traditional detection methods by incorporating real world human knowledge. Dadvar et al. [39] also adopted a mixed-initiative approach to cyberbullying detection by using a panel of cyberbullying experts to provide weighting to features set of user-based information such as the age of the user, membership duration, the number of uploads, the number of subscriptions, the total number of posts, and length of the post. The human experts rated each feature on its relative importance and the likelihood that a bully can be identified by the feature.

Proposed Model

Ensemble techniques are the methods that use multiple learning algorithms or models to produce one optimal predictive model. The model produced has better performance than the base learners taken alone. The proposed system has developed various ensemble models out of which optimal result will be provided by proposed ensemble model. The general ensemble framework is shown below:



The figure shown above represents working of an ensemble model. Stacking often considers heterogeneous weak learners, learns them in parallel, and combines them by training a meta-learner to output a prediction based on the different weak learner's predictions. In the proposed work we have developed an optimized ensemble model to produce good results.

3.1 Detailed of Proposed system

The purpose of the proposed system is to classify offensive text using optimized ensemble model. The proposed system has develop an ensemble model consists of a combination of Logistic Regression, Decision Tree, Random Forest and SGD Classifiers. To find an optimal ensemble model, the proposed system has uses following base learners:

- Logistic regression (LR)
- Decision tree (DT)
- Random forest (RF)
- Stochastic gradient descendent (SGD)
- K-nearest neighbour (KNN)
- Multinomial naïve bayes (MNB)
- Support vector machine (SVM)
- AdaBoost

Then the proposed method has developed many ensemble models by using these base learners.

3.2 Proposed Model algorithm

Proposed Algorithm Ensemble_of_Classifiers ()
{
Step 1: Read the dataset of hate tweets.

Step 2: Clean the data.
Step 3: Split the dataset into train and test set.
Step 4: Create Ensemble models using different combinations of Base Classifiers'.
Step 5: Train the models using various ensemble classifiers.
Step 6: Evaluate the models.
Step 7: Compare the results.
Step 8: End of algorithm.
}

III. RESULTS

Evaluations of various classifier algorithms according to accuracy are displayed below.

Table 6.1: Performance Evaluation.

Method	Testing Accuracy on dataset1 (%)
Logistic Regression	**69.72**
Random Forest	**69.72**
AdaBoost	**70.68**
SGD	**66.78**
KNN	**53.23**
Decision Tree	**65.02**
Multinomial Naive Bayes	**66.86**
Ensemble1	**73.54**
Ensemble2	**73.02**
Ensemble3	**71.05**
Ensemble4	**72.29**
Ensemble5	**73.25**
Ensemble6	**72.41**
Ensemble7	**73.07**
Ensemble8	**72.21**

It is observed that proposed classifier ensemble1 which is a combination of Logistic Regression, Decision Tree, Random Forest and SGD Classifiers, gives the better results in terms of accuracy.

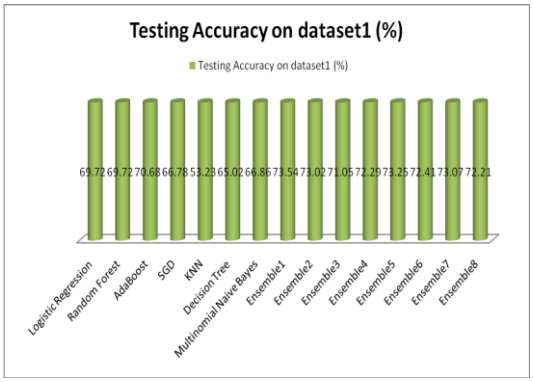


Figure 6.47: Accuracy Chart of all the models.

IV. CONCLUSION

This work attempted to maximize the models' performance by improving feature engineering and ML model building. The first solution tried to fine-tune the pre-trained models with new data from the training dataset; thus, the unseen vocabularies were added to the model. The results showed a slight improvement in the performance, and this was because the training dataset size that we used to retrain the word embedding models.

Even though the previous model resulted in powerful algorithms such as KNN, DT and SVM, the results were not convincing. This was due to the problem that we could not discover the best hyper parameters of the two models that generated the maximum accuracy. Therefore, we used a hybrid approach of classifiers in the proposed phase. The results showed a significant improvement in the models' performance.

I References

[1] S. Z. Omar, A. Daud, M. S. Hassan, J. Bolong, and M. Teimmouri, "Children internet usage: Opportunities for self development," Procedia Social Behav. Sci., vol. 155, pp. 75_80, Nov. 2014.
[2] L. Haddon and S. Livingstone, ``Risks, opportunities, and risky opportunities: How children make sense of the online environment," in Cognitive Development in Digital Contexts. Amsterdam, The Netherlands: Elsevier, 2017, pp. 275_302.
[3] T. K. H. Chan, C. M. K. Cheung, and R. Y. M.Wong, "Cyberbullying on social networking sites: The crime opportunity and affordance perspectives," J. Manage. Inf. Syst., vol. 36, no. 2, pp. 574_609, Apr. 2019.
[4] M. Duggan, ``Online harassment 2017," Pew Res. Centre, Washington, DC, USA, Tech. Rep., Jul. 2017.
[5] G. M. Abaido, "Cyberbullying on social media platforms among university students in the United Arab Emirates," Int. J. Adolescence Youth, vol. 25, no. 1, pp. 407_420, Dec. 2020, doi: 10.1080/02673843. 2019.1669059.
[6] F. Sticca, S. Ruggieri, F. Alsaker, and S. Perren, "Longitudinal risk factors for cyberbullying in adolescence," J. Community Appl. Social Psychol., vol. 23, no. 1, pp. 52_67, Jan. 2013.
[7] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in Proc. ACM Conf. Web Sci. (WebSci). New York, NY, USA: Association for Computing Machinery, 2017, pp. 13_22.
[8] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning," in Proc.

IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, Jul. 2017, pp. 409_416.

[9] M. Dadvar, D. Trieschnigg, and F. de Jong, “Experts and machines against bullies: A hybrid approach to detect cyber bullies," in Proc. 27th Can. Conf. Artif. Intell. Adv. Artif. Intell. Can. (AI) in Lecture Notes in Computer Science, vol. 8436, M. Sokolova and P. van Beek, Eds. Montreal, QC, Canada: Springer, May 2014, pp. 275_281, doi: 10.1007/978-3-319-06483-3_25.

[10] K. Dinakar, R. Reichart, and H. Lieberman, “ Modelling the detection of textual cyberbullying," in Proc. Social Mobile Web, Papers (ICWSM) Workshop, Barcelona, Spain, Jul. 2011, pp. 1_7.

[11] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying," in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA), vol. 2, Dec. 2011, pp. 241_244, doi:10.1109/ICMLA.2011.152.

[12] A. Kumar, S. Nayak, and N. Chandra, “Empirical analysis of supervised machine learning techniques for cyberbullying detection," in Proc. Int. Conf. Innov. Comput. Commun., S. Bhattacharyya, A. E. Hassanien, D. Gupta, A. Khanna, and I. Pan, Eds. Singapore: Springer, 2019, pp. 223_230.

[13] S. Agrawal and A. Awekar, “Deep learning for detecting cyberbullying across multiple social media platforms," in Advances in Information Retrieval, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham, Switzerland: Springer, 2018, pp. 141_153.

[14] T. Mikolov, M. Karafiat, L. Burget, J. cernocky, and S. Khudanpur, “Recurrent neural network based language model," in Proc. 11th Annu. Conf. Int. Speech Commun. Assoc., vol. 2, 2010, pp. 1045_1048.

[15] C. Chelmis, D.-S. Zois, and M. Yao, “Mining patterns of cyberbullying on Twitter," in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Nov. 2017, pp. 126_133.

[16] L. Cheng, J. Li, Y. Silva, D. Hall, and H. Liu, ``PI-bully: Personalized cyberbullying detection with peer influence," in Proc. 28th Int. Joint Conf. Artif. Intell. Aug. 2019, pp. 5829_5835.

[17] B. Belsey, “Cyberbullying: An emerging threat to the `always on' generation," Recuperado el, vol. 5, no. 5, p. 2010, and 2005.

[18] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and H. H. Reese, “Cyber bullying among college students: Evidence from multiple domains of college life," in Misbehaviors Online in Higher Education. Bingley, U.K.: Emerald Group Publishing Limited, 2012.

[19] J. W. Patchin and S. Hinduja, Cyberbullying Prevention and Response*:* Expert Perspectives. Evanston, IL, USA: Routledge, 2012.

[20] P. K. Smith, “Cyberbullying and cyber aggression," in Handbook of School Violence and School Safety. Evanston, IL, USA: Routledge, 2012, pp. 111_121.

[21] M. Mladenovic, V. O²mjanski, and S. V. Stankovic, “Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges," ACM Comput. Surv. vol. 54, no. 1, pp. 1_42, Apr. 2021, doi: 10.1145/3424246.

[22] N. Tahmasbi and A. Fuchsberger, “Challenges and future directions of automated cyberbullying detection," Amer. Conf. Inf. Syst., USA, Tech. Rep. 9780996683166, 2018.

[23] X. Tian, “Investigating cyberbullying in social media: The case of Twitter," in *Proc.* KSU Conf. Cyber secure Educ., Res. Pract., Atlanta, GA, USA, 2016.

[24] T. Mahlangu, C. Tu, and P. Owolawi, “A review of automated detection methods for cyberbullying," in Proc. Int. Conf. Intell. Innov. Comput. Appl. *(*ICONIC), Dec. 2018, pp. 1_5.

[25] Z.Waseem and D. Hovy, “Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. NAACL Student Res. Workshop, 2016, pp. 88_93, doi: 10.18653/v1/n16-2013.

[26] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on twitter using a convolution-GRU based deep neural network," in The Semantic Web, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham, Switzerland: Springer, 2018, pp. 745_760.

[27] A. Kumar, S. Nayak, and N. Chandra, “Empirical analysis of supervised machine learning techniques for cyberbullying detection," in *Proc. Int.* Conf. Innov. Comput. Commun., S. Bhattacharyya, A. E. Hassanien, D. Gupta, A. Khanna, and I. Pan, Eds. Singapore: Springer, 2019, pp. 223_230.

[28] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text," ACM Comput. Surv. vol. 51, no. 4, pp. 1_30, Sep. 2018.

[29] International Journal of Advanced Computer Science and Applications,(IJACSA) Vol. 9, No. 9, 2018 - Deep Learning Algorithm for Cyberbullying Detection by Monirah Abdullah Al-Ajlan and Mourad Ykhlef, University of Riyadh, Saudi Arabia.

[30] A. Alakrot, M. Fraifer and N. S. Nikolov, "Machine learning approach to detection of offensive language in online communication in Arabic", Proc. IEEE 1st Int. Maghreb Meeting Conf. Sci. Techn. Autom. Control Comput. Eng. (MI-STA), pp. 244-249, May 2021.

[31] H. Karayigit, Ç. I. Aci and A. Akdagli, "Detecting abusive Instagram comments in Turkish using convolutional neural network and machine learning methods", Expert Syst. Appl., vol. 174, Jul. 2021.

[32] A. Ziani, N. Azizi, D. Zenakhra, S. Cheriguene and M. Aldwairi, "Combining RSS-SVM with genetic algorithm for Arabic opinions analysis", Int. J. Intell. Syst. Technol. Appl., vol. 18, no. 1, pp. 152-178, 2019.

[33] Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. (2012a). Common Sense Reasoning For Detection, Prevention, and Mitigation of Cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3).

[34] Perez, P.J.C., Valdez, C.J.L., Ortiz, M.D.G.C., Barrera, J.P.S. and Perez, P.F. MISAAC: Instant Messaging Tool for Cyberbullying Detection [online]. Available from http://worldcomp-proceedings.com/proc/p2012/ICA7994.pdf [Accessed 21st June 2015].

[35] Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). Detecting Cyberbullying: Query Terms and Techniques. IN: Annual ACM Web Science Conference. 5th. Indiana. June 23 – 26, 2013. New York: ACM, 195-204.

[36] Nahar, V., Al-Maskari, S., Li, X. and Pang, C. (2014). "Semi supervised Learning for Cyberbullying Detection in Social Networks", Databases Theory and Applications, 8506, p.160-171.

[37] Bretschneider, U., Wohner, T., and Peters, R. (2014). Detecting Online Harassment in Social Networks [online]. Available from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1003&context =icis2014.

[38] Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S. and Mattson, S.A. (2015). Careful what you share in six seconds: detecting cyberbullying instances in Vine. IN: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris. August 25-28, 2015. ACM, 617-622.

[39] Dadvar, M., De Jong, F.M.G., Ordelman, R. J. F. and Trieschnigg, R. B. (2012a). Improved Cyberbullying Detection Using Gender Information [online]. Available from http://eprints.eemcs.utwente.nl/21608/01/DIR12_reviewed04.pdf.

[40] Dinakar, K., Reichart, R. and Lieberman, H. (2011), "Modelling the Detection of Textual Cyberbullying. The Social Mobile Web.

[41] Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A. and Edwards, L. (2009). Detection of Harassment on Web 2.0. IN: Content Analysis in the WEB. Madrid. April 21, 2009.

[42] Xu, J.M., Jun, K.S., Zhu, X. and Bellmore, A. (2012a). Learning from Bullying Traces in Social Media. IN: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montreal. June 3 – 8, 2012. Stroudsburg: ACL, 656-666.

[43] Sood, S.O., Antin, J. and Churchill, E.F. (2012a). Using Crowd sourcing to Improve Profanity Detection. IN: AAAI Spring Symposium: Wisdom of the Crowd. Stanford, March 26 – 28, 2012. Palo Alto: The AAAI Press, 69 – 74.

[44] Munezero, M., Montero, C.S., Kekkonen, T., Sutinen, E., Mozgovoy, M. and Klyuev, V. (2014). Automatic Detection of Antisocial Behaviour in Texts. Informatics. Special Issue: Advances in Semantic Information Retrieval, 38(1), p.3 – 10.