Social Media Cyberbullying Detection using Machine Learning

John Hani[1], Mohamed Nashaat[2], Mostafa Ahmed[3],
Zeyad Emad[4], Eslam Amer[5]
Department of Computer Science,
Misr International University, Cairo, Egypt

Ammar Mohammed[6]
Department of Computer Science, ISSR
Cairo University, Cairo, Egypt

***Abstract*—With the exponential increase of social media users, cyberbullying has been emerged as a form of bullying through electronic messages. Social networks provides a rich environment for bullies to uses these networks as vulnerable to attacks against victims. Given the consequences of cyberbullying on victims, it is necessary to find suitable actions to detect and prevent it. Machine learning can be helpful to detect language patterns of the bullies and hence can generate a model to automatically detect cyberbullying actions. This paper proposes a supervised machine learning approach for detecting and preventing cyberbullying. Several classifiers are used to train and recognize bullying actions. The evaluation of the proposed approach on cyberbullying dataset shows that Neural Network performs better and achieves accuracy of 92.8% and SVM achieves 90.3. Also, NN outperforms other classifiers of similar work on the same dataset.**

***Keywords*—Cyberbullying; machine learning; neural network**

I. Introduction

With the increasing number of users on social media leads to a new way of bullying. The later term, is defined as an intentional or an aggressive acts which are carried out by person or groups of individuals using repeatedly communication messages over time against a victim who cannot easily defend him or herself [1]. Bullying has always been a part of society. With the inception of the internet, it was only a matter of time until bullies found their way on to this new and opportunistic medium. Using services like email and instant messenger, bullies became able to do their nasty deeds with anonymity and great distance between them and their targets. According to Cambridge dictionary the term cyberbullying is defined as the activity of using the internet to harm or frighten another person, especially by sending them unpleasant messages. The main factor that separates cyberbullying from traditional bullying is the effect that it has on the victim. Traditional bullying may end in physical damage as well as emotional and psychological damage, as opposed to cyberbullying, where it is all emotional and psychological.

Given the consequences of cyberbullying on victims, it is urgently needed to find a proper actions to detect and hence to prevent it. One of the successful approaches that learns from data and generates a model that automatically classifies proper actions is machine learning. Machine learning can be helpful to detect language patterns of the bullies and hence can generate a model to detect cyberbullying actions. Thus, the main contribution of this paper is to propose a supervised machine learning approach for detecting and preventing cyberbullying. The proposed approach is evaluated on a cyberbullying dataset from kaggle which was collected and labeled by the authors Kelly Reynolds et al. in their paper [2]. The performance of SVM and Neural Network classifiers are compared on both TFIDF and sentiment analysis feature extraction methods. Furthermore, experiments were made on different n-gram language model. 2-gram, 3-gram and 4-gram has been taken into consideration during the evaluation of the model produced by the classifiers. Finally, we evaluate our proposed approach with previous related work who used the same data.

The rest of the paper is organized as follows. Section II shows several related work. Section III describes the proposed approach. Section IV shows the experimental results and the evaluation of the proposed approach. Finally, Section V concludes the paper.

II. Related Work

There are many approaches that proposes systems which can detect cyberbullying automatically with high accuracy. First one is author Nandhini et al. [3] have proposed a model that uses Naïve Bayes machine learning approach and by their work they achieved 91% accuracy and got their dataset from MySpace.com, and then they proposed another model [4] Naïve Bayes classifier and genetic operations (FuzGen) and they achieved 87% accuracy. Another approach by Romsaiyud et al. [5] they enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering and by this approach they achieved 95.79% accuracy on datasets from Slashdot, Kongregate, and MySpace. However, they have a problem that the cluster processes doesn't work in parallel. Moreover, in the approach proposed by Bunchanan et al. [6] they used War of Tanks game chat to get their dataset and manually classified them and then compared them to simple Naïve classification that uses sentiment analysis as a feature, their results were poor when compared to the manually classified results. Furthermore, Isa et al. [7] proposed an approach after getting their dataset from kaggle they used two classifier Naïve Bayes and SVM. The Naïve Bayes classifier yielded average accuracy of 92.81% while SVM with poly kernel yielded accuracy of 97.11%, but they did not mention their training or testing size of the dataset, so the results may not be credible. Another Approach by Dinakar et al. [8] that aimed to detect explicit bullying language pertaining to (1) Sexuality, (2) Race & Culture and (3) intelligence, they acquired their dataset from YouTube comment section. After applying SVM and Naïve Bayes classifiers, SVM yielded accuracy of 66%

and Naïve Bayes 63%. Moving on to Di Capua et al. [9], they proposed a new way for cyberbullying detection by adopting an unsupervised approach, they used the classifiers inconsistently over their dataset, applying SVM on FormSpring and achieving 67% on recall, applying GHSOM on YouTube and achieving 60% precision, 69% accuracy and 94% recall, applying Naïve Bayes on Twitter and achieving 67% accuracy. Additionally, Haidar et al. [10] proposed a model to detect cyberbullying but using Arabic language they used Naïve Bayes and achieved 90.85% precision and SVM achieved 94.1% as precision but they have high rate of false positive also the are work on Arabic language.

Another type of approaches using Deep Learning and Neural Networks. One of the proposed methods is Zhang et al. [11] in their paper uses novel pronunciation based convolution neural network (PCNN), thereby alleviating the problem of noise and bullying data sparsity to overcome class imbalance. 1313 messages from twitter, 13,000 messages from formspring.me. Accuracy of the twitter dataset wasn't calculated due to it being imbalanced. While Achieving 56% on precision, 78% recall and 96% accuracy, while achieving high accuracy their dataset was unbalanced, so that gives false results and that reflects in precision score which is 56%. The authors Nobata et al. [12] showed that using abusive language has increased recently, They used a framework called Vowpal wabbit for classification, and they also developed a supervised classification methodology with NLP features that outperform deep learning approach, The F-Score reached 0.817 using dataset collected from comments posted on Yahoo News and Finance.

Zhao et al. [13] proposed framework specific for cyberbullying detection, they used word embedding that makes a list of pre-defined insulting words and assign weights to obtain bullying features, they used SVM as their main classifier and got recall of 79.4%. Then another approach was proposed by Parime et al. [14] they got their dataset from MySpace and manually marked them and they used the SVM Classifier for the classification. Moreover, Chen et al. [15] proposed a new feature extraction method called Lexical Syntactic Feature and SVM as their classifier and they achieved 77.9% precision and 77.8% recall. Furthermore, Ting et al. [16] proposed a technique based on SNM, they collected their data from social media and then used SNA measurements and sentiments as features. Seven experiments were made and they achieved around 97% precision and 71% as recall. Furthermore, Harsh Dani et al. [17] introduced a new framework called SICD, they used KNN for classification. Finally, they achieved 0.6105 F1 score and 0.7539 AUC score.

SVM classifier was one of the approaches used in the research papers. Dadvar et al. [18][19][20][21] have constructed in the first and second paper a Support Vector Machine classifier using WEKA, their dataset was collected from Myspace. They achieved 43% on precision, 16% in recall and they didn't mention the accuracy, the only difference between the two papers is that they used gender information in classification in the second paper. Moreover, in their second paper 4626 comments from 3858 distinct users were collected. The comments were manually labelled as bullying (9.7%) and non-bullying (inter-annotator agreement 93%). SVM classifier was applied by them and were able to reach results of up to 78% on precision and 55% on recall. Finally, in their third paper they applied 3 models for their dataset gathered from YouTube comment section: Multi-Criteria Evaluation Systems (MCES), machine learning: (Naïve Bayes classifier, decision tree, SVM), Hybrid approach. The MCES score 72% on accuracy, while Naïve Bayes scored the highest out of the three with 66%. Moving on to another author, Potha et al. [22] have also used the SVM approach and achieved 49.8% result on accuracy. While Chavan et al. [23] used two classifiers: logistic regression and support vector machine. The logistic regression achieved 73.76 accuracy and 60% recall and 64.4% Precision. While for the support vector machine they achieved 77.65% accuracy and 58% recall and 70% precision's and they got their dataset from Kaggle.

III. Proposed Approach

The proposed approach, as seen in Fig. 1, contains three main steps: Preprocessing, features extraction and classification step. In the preprocessing step we clean the data by removing the noise and unnecessary text. The preprocessing step is done in the following:

- Tokenization: In this part we take the text as sentences or whole paragraphs and then output the entered text as separated words in a list.
- Lowering text: This takes the list of words that got out of the tokenization and then lower all the letters Like: 'THIS IS AWESOME' is going to be 'this is awesome'.
- Stop words and encoding cleaning: This is an essential part of the preprocessing where we clean the text from those stop words and encoding characters like \n or \t which do not provide a meaningful information to the classifiers.
- Word Correction: In this part we used Microsoft Bing word correction API [24] that takes a word and then return a JSON object with the most similar words and the distance between these words and the original word.

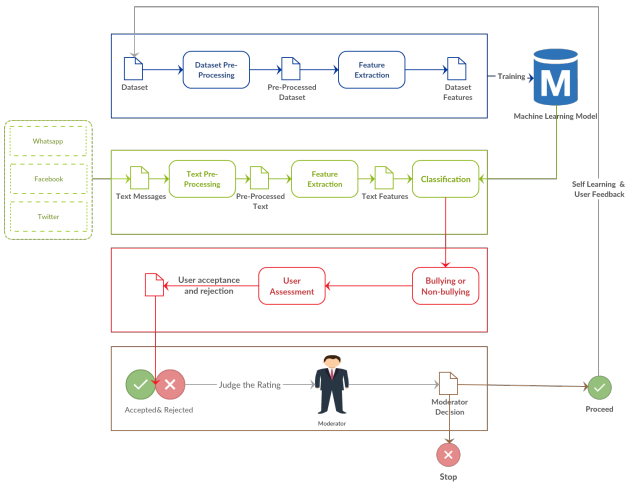

Fig. 1. Proposed Approach

The second step of the proposed Model is the features extraction step. In this step the textual data is transformed into a suitable format applicable to feed into machine learning algorithms. First we extract the features of the input data using TFIDF[25] as and put them in a features list. The key idea of TFIDF is that it works on the text and get the weights of the words with respect to the document or sentence. In Addition to TFIDF, we use sentiment analysis technique[26] to extract the polarity of the sentences and add them as a feature into the features list containing the TFIDF features. The polarity of the sentences means that if the sentence is classified as positive or negative. For that purpose we extract the polarity using Text Blob library[27] which is a pre-trained model on movie reviews. In addition to the feature extraction using TFIDF and sentiment polarity extraction, the propose approach uses N-Gram[28] to consider the different combinations of the words during evaluation of the model. Particularly, we use used 2-Gram, 3-Gram and 4-Gram.

The last step in the proposed approach is the classification step where the extracted features are fed into a classification algorithm to train, and test the classifier and hence use it in the prediction phase. We used two classifiers, namely, SVM (Support Vector Machine) and Neural Network. The neural network contains three layers: Input, hidden, output layer. In the input layer, it consists of 128 nodes. In the hidden layer, it contains 64 neurons. The output layer is a Boolean output.

Generally, the evaluation of classifiers is done using several evaluation matrices depends on the confusion matrix. Among of those criteria are Accuracy, precision, recall and f-score. They are calculated according to the following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F-Score = \frac{2*precision*recall}{precision+recall} \quad (4)$$

Where TP represents the number of true positive, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives classes.

IV. Experimental Results

This section describes the experimental results on the proposed approach. We evaluate the proposed approach on the cyberbullying dataset from kaggle. In the following we describes the Data and the results.

A. Data Description

We have used cyberbullying dataset from Kaggle which was collected and labeled by the authors Kelly Reynolds et al. in their paper [2]. This dataset contains in general 12773 conversations messages collected from Formspring. The dataset contains questions and their answers annotated with either cyberbullying or not. The annotation classes were unbalanced distributed such that 1038 question-answering instances out of 12773 belongs to the class cyberbullying, while 11735 belongs to the other class. First, to remedy the data unbalancing, we take the same number instances of both classes to measure the accuracy. We also removed from the data big size conversations and remove the noisy data. We ended up with total 1608 instance conversations where 804 instances belongs to each class. Table I summarizes the statistics of dataset.

TABLE I. Statistics of the Dataset

Total number of Conversations	1608
Number of cyberbullying	804
Number of non-Cyberbullying	804
Number of distinct words	5628
Number of token	48843
Maximum Conversation size	773 Characters
Minimum Conversation size	59 Characters

B. Results

After preprocessing the dataset, we follow the same step presented in Section III to extract the features. We then split the dataset into ratios (0.8,0.2) for train and test. Accuracy, recall and precision, and f-score are taken as a performance measure to evaluate the classifiers. We apply SVM as well as Neural Network (NN) as they are among the best performance classifiers in the literature. We run several experiments on different n-gram language model. In Particular, we take into consideration 2-gram, 3-gram, and 4-gram during the evaluation of the model produced by the classifiers. Table II summarizes the accuracy of both SVM and NN. The SVM classifier achieved the highest percentage using 4-Gram with accuracy 90.3% while the NN achieved highest accuracy using 3-Gram with accuracy 92.8%. It is found that the average accuracy of all n-gram models of NN achieves 91.76%, while the average accuracy of all n-gram models of SVM achieves 89.87%. Fig. 2 depicts the accuracy results of both classifiers.

TABLE II. The Accuracy of SVM and NN in Different Language Model

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.9%	**90.3%**	89.87%
Neural Network	90.9%	**92.8%**	91.6%	91.76%

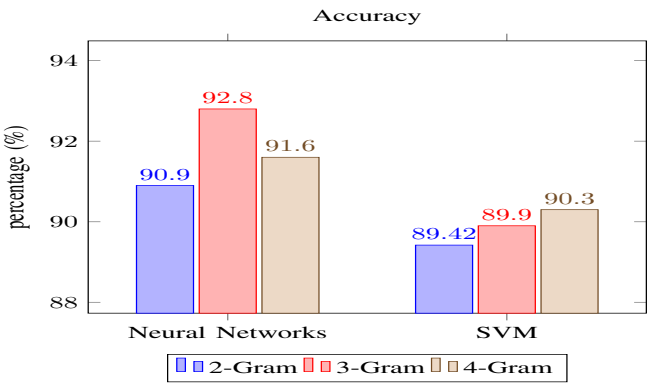

Fig. 2. Comparison between SVM and Neural Network in Terms of Accuracy

In addition to accuracy, Table III and Table IV show the evaluations of both classifiers in terms of precision and recall respectively for each language model. The trade-off between recall and precision is shown in Table V which represents the f-score of both classifiers in the different language model. Table V summarizes the f-score of both SVM and NN. The SVM classifier achieved the highest f-measure using 4-Gram with f-score 90.3% while the NN achieved highest f-measure using 2-Gram with f-score 92.2%. It is found that the average f-score of all n-gram models of NN achieves 91.9%, while the average f-score of all n-gram models of SVM achieves 89.8%. Fig. 3 summarizes the f-score of the classification of the SVM and Neural Network. The results of average accuracy as well as the average f-score indicate that NN performs better than SVM.

TABLE III. RECALL OF SVM AND NN

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	90.3%	**90.8%**	90.1%
Neural Network	91.6%	91.5%	**92%**	91.7%

TABLE IV. PRECISION OF SVM AND NN

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.5%	**90%**	89.6%
Neural Network	**93%**	92.5%	91.7%	92.4%

TABLE V. F-SCORE OF SVM AND NN

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.8%	**90.3%**	89.8%
Neural Network	**92.2%**	91.9%	91.8%	91.9%

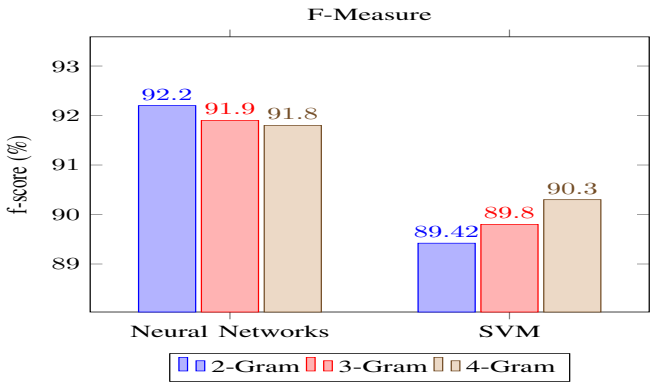

Fig. 3. Comparison between SVM and Neural Network in Terms of F-Measure

In addition to the previous experiments, we evaluate and compare our classifiers on the proposed approach with the work of [23]. In this work, they used logistic regression and SVM for classification and used the same data. Moreover, we have calculated the average accuracy, recall, precision and F-score of our two classifiers. The summary of results is shown in Table VI. To compare the work, it is found that our proposed NN model outperforms all other classifiers and is ranked as the best results in terms of average accuracy and F-Score achieving accuracy 91.76% and f-score 91.9%. In Fig. 4 we are comparing between our best classifier with their best classifier in case of accuracy. Finally, here in Fig. 5 we are comparing between our best classifier with their best classifier in case of F-Measure.

TABLE VI. COMPARISON WITH RELATED WORK

	Classifier	Avg. Accuracy	Avg. Recall	Avg. Precision	Avg. F-Score
Vikas S Chavan	Logistic regression	73.76	61.47%	64.4%	62.9%
	SVM	77.65%	58.29%	70.29%	63.7%
Current Results	Neural Network	**91.76%**	**91.7%**	**92.4%**	**91.9%**
	SVM	89.87%	90.1%	89.6%	89.8%

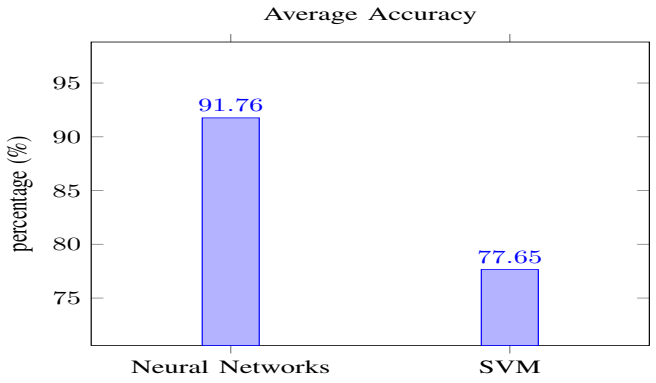

Fig. 4. Comparison between the Best Classifiers in Terms of Accuracy

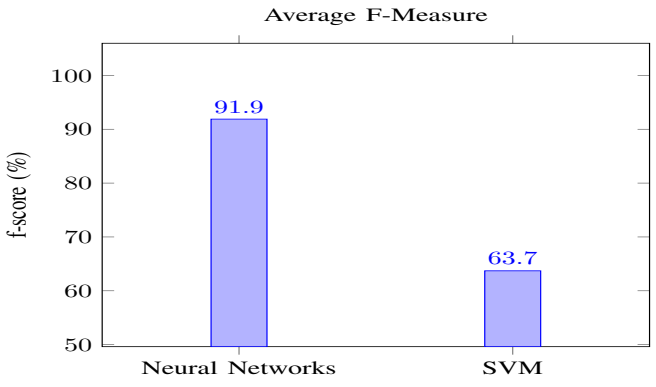

Fig. 5. Comparison between the Best Classifiers in Terms of F-Measure

V. Conclusion

In this paper, we proposed an approach to detect cyberbullying using machine learning techniques. We evaluated our model on two classifiers SVM and Neural Network and we used TFIDF and sentiment analysis algorithms for features extraction. The classifications were evaluated on different n-gram language models. We achieved 92.8% accuracy using Neural Network with 3-grams and 90.3% accuracy using SVM with 4-grams while using both TFIDF and sentiment analysis together. We found that our Neural Network performed better than the SVM classifier as it also achieves average f-score 91.9% while the SVM achieves average f-score 89.8%. Furthermore, we compared our work with another related work that used the same dataset, finding that our Neural Network outperformed their classifiers in terms of accuracy and f-score. By achieving this accuracy, our work is definitely going to improve cyberbullying detection to help people to use social media safely. However, detecting cyberbullying pattern is limited by the size of training data. Thus, a larger cyberbullying data is needed to improve the performance. Hence, deep learning techniques will be suitable in the larger data as they are proven to outperform machine learning approaches over larger size data.

References

[1] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385, 2008.

[2] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE, 2011.

[3] B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, page 20. ACM, 2015.

[4] B Sri Nandhini and JI Sheeba. Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45:485–492, 2015.

[5] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. Automated cyberbullying detection using clustering appearance patterns. In *Knowledge and Smart Technology (KST), 2017 9th International Conference on*, pages 242–247. IEEE, 2017.

[6] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security*, 76:197–213, 2018.

[7] Sani Muhamad Isa, Livia Ashianti, et al. Cyberbullying classification using text mining. In *Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on*, pages 241–246. IEEE, 2017.

[8] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[9] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised cyber bullying detection in social networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 432–437. IEEE, 2016.

[10] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Advances in Science, Technology and Engineering Systems Journal*, 2(6):275–284, 2017.

[11] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 740–745. IEEE, 2016.

[12] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[13] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, page 43. ACM, 2016.

[14] Sourabh Parime and Vaibhav Suri. Cyberbullying detection and prevention: Data mining and psychological perspective. In *Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on*, pages 1541–1547. IEEE, 2014.

[15] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[16] I-Hsien Ting, Wun Sheng Liou, Dario Liberona, Shyue-Liang Wang, and Giovanny Mauricio Tarazona Bermudez. Towards the detection of cyberbullying based on social network mining techniques. In *Behavioral, Economic, Socio-cultural Computing (BESC), 2017 International Conference on*, pages 1–2. IEEE, 2017.

[17] Harsh Dani, Jundong Li, and Huan Liu. Sentiment informed cyberbullying detection in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–67. Springer, 2017.

[18] Maral Dadvar and Franciska De Jong. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*, pages 121–126. ACM, 2012.

[19] Maral Dadvar, de FMG Jong, Roeland Ordelman, and Dolf Trieschnigg. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent, 2012.

[20] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[21] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pages 275–281. Springer, 2014.

[22] Nektaria Potha and Manolis Maragoudakis. Cyberbullying detection using time series modeling. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 373–382. IEEE, 2014.

[23] Vikas S Chavan and SS Shylaja. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *Advances in computing, communications and informatics (ICACCI), 2015 International Conference on*, pages 2354–2358. IEEE, 2015.

[24] Youssef Bassil and Mohammad Alwani. Post-editing error correction algorithm for speech recognition using bing spelling suggestion. *arXiv preprint arXiv:1203.5255*, 2012.

[25] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[26] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[27] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 2014.

[28] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer, 1994.