RESEARCH ARTICLE OPEN ACCESS

Survey On Cyberbullying Detection on Social Networks Using DL Approaches

Abhirenjini K A [1], Candiz Rozario [2], Kripa Treasa [3], Juvariya Yoosuf [4], Vidya Hari [5]

[1], [2], [3], [4] UG Scholor,Dept.Of Computer Science and Engineering KMEA Engineering College, Kerala-India (of Aff.KTU) Thrissur, India

[5] Asst.Prof,Dept.Of Computer Science and Engineering KMEA Engineering College, Kerala-India (of Aff.KTU)Ernakulam, India

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 sometimes even forces 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 texts or messages on social media has gained a growing amount of attention among researchers. This research aims to design and develop an effective technique to detect online abusive and bullying messages by merging convolutional neural networks (CNN) and deep learning. Six distinct features, namely bagof-words (BoW) and term frequency-inverse text frequency (TFIDF), ngrams, sentiment scores, finding profanity words, and counting emojis, are used to analyse the accuracy level of the deep learning algorithm.

Keywords: - CNN,Abuse and crime involving computers, deep learning, sentiment analysis, social networking

I. INTRODUCTION

Due to the increasing usage of text, image, and video-based communication in social media applications, cyberbullying events are growing. These events need to be detected and prevented before causing harm to users. The data was extracted from GitHub and Kaggle and tested with Telegram realtime data. The model was developed with a combination of convolutional neural networks (CNN) and long-short-term memory (LSTM) using Kears functional API. This model predicted more accurately; the image-based prediction gives 86% accuracy, and the text-based prediction gives 85% accuracy after training the model. An effective user interface system will be developed to prevent cyberbullying.

New definitions of social communication have emerged as a result of the development of social networks and the internet in the digital age. Social media platforms including Telegram, Facebook, WhatsApp, Instagram, and Snapchat are used by 58.11% of people worldwide to communicate, and more cases of cyberbullying have been documented in recent years. A group or a person can intentionally cause harm by engaging in cyberbullying. Additionally, it occasionally grows. Isolation, rage, the psychological impacts of sadness and anxiety, scholastic problems, suicidal thoughts, self-harm, the behavioural implications of drug or alcohol use, and skipping normal activities are some of the repercussions brought on by cyberbullying. The use of emoticons, memes, text, and characters in images in communication makes it difficult to spot cyberbullying. Bullying incidents are now a wellorganized and multi-media source of data. Websites for social networking are just concerned with photo sharing. In order to urge victims to engage in cyberbullying, these trends switch from text to images.

The cyberbullying content is classified as bully or non-bully. Consider the case where we need to identify cyberbullying on a social media site such as Twitter. The text of tweets would be analysed by a text-based model to spot any words or phrases that suggest harassment, aggression, or discrimination. Tweets containing the words ”kill yourself”, ”ugly” or ”stupid”, for instance, might be labelled as cyberbullying. Again, consider that we want to identify cyberbullying on an Instagramlike photo-sharing app. Images with offensive gestures, hate symbols, or violent scenes are just a few examples of visual cues that can be used by an image-based model to analyse the content of images and detect cyberbullying.

The purpose of this research is to design and develop an effective technique to detect online abusive and bullying messages by merging convolutional neural networks (CNN) and deep learning. There are six distinct features, namely, bag-of-words (BoW), term frequency, inverted text frequency (TFIDF), Ngrams sentiment scores, finding profanity words, and counting emoji. The datasets are collected from Kaggle, which consists of toxic Twitter comments, YouTube comments, and comments from Formspring. The dataset is preprocessed and then vectorized with TF-IDF, n-gramme, bag of words (BOW), sentiment scores, finding profanity words, and counting emojis. Then split the dataset into training and testing sets. Now, these datasets are fed into the three models, namely Perceptron, LR, and SVM, and after all these processes, the three models are ensembled into the voting classifier. The dataset, which is processed and categorised as bullying and non-bullying, The features like finding profinity words, emojis, and bags of words have been extracted from the tweets. The text of tweets would be analysed to spot any words or phrases that suggest harassment, aggression, or discrimination.

II. LITERATURE SURVEY

This section briefly discusses a few notable review papers on machine learning-based cyberbullying detection. We also present a comparison between our work with these existing works to show the novelty of our work.

Yin et al. [8] carried out the first research into the automatic recognition of online cyberbullying. The authors used three different datasets to detect harassment on three different online platforms. The Kongregate platform was used for one dataset collection, while the other datasets were gathered from discussion-based communities (e.g., Reddit). A linear kernel classification model and various feature extraction methods (N-grammes and term frequency-inverse term frequency (TFIDF)) were employed for the classification task. Although their experimental results were ambiguous, the study served as a starting point for further investigation. Another study was proposed in the same field by [9]. The authors implemented C4.5, k-nearest neighbours (KNN), and support vector machine (SVM) classification techniques, which were tested on a dataset consisting of text comments collected from the Formspring platform. Based on their experimental results, the C4.5 decision tree algorithm surpassed both the KNN and SVM classifiers with a detection rate accuracy of 78.5%. Dinakar et al. [11] proposed a two-step detection method. The very first step was to decide whether or not a piece of information or content falls under the category of sensitive. The second step involved classifying the content of the text with a particular label (e.g., intellectual ability or sexual orientation). The proposed method was tested on 4500 YouTube comments, and the classification accuracy ranged from 70 to 80%.

Dadvar et al. [12] proposed a gender-based method to detect cyberbullying related to gender harassment. Their approach employed two distinct vocabulary sets. Based on their findings, this method has marginally enhanced the accuracy of machinelearning classifiers. Subsequently, several studies using a range of different techniques have been conducted in relation to cyberbullying detection. Based on Essential Dimensions of Latent Semantic Indexing (EDLSI), Kontostathis et al. [13] developed a model for classifying the most popular words used in cyberbullying based on messages from the Formspring.me website. The authors reported that the classification model provided an average precision of 91.25%. Ptaszynski [14] used brute force search algorithms and learning classifiers to find patterns associated with online cyberbullying. Specifically, in their classification process they extracted patterns from sentences. Based on the Human Rights Center database, this approach surpassed earlier cyberbullying detection methodologies.

Zhang et al. [15] used deep learning to design a robust cyberbullying identification model. A convolutional neural network (CNN) model was built using the pronunciation of words as input features for the detection process. The CNN model was tested and verified on a dataset consisting of social media text comments gathered from the Twitter and Formspring.me platforms. The results showed that the pronunciation-based CNN model performed better than baseline CNN models with arbitrarily created word embeddings. Chavan and Shyla [16] presented a method for determining whether a comment would be insulting to other users. They used skip-grams as input sequences for their machine learning classifiers. Furthermore, they incorporated the results of SVM and logistic regression classification models into their methodology. Squicciarini et al. [17] used a decision tree classifier to identify text-based features and then presented a rule-based method to further identify cyberbullying behaviors.

According to the literature review on cyberbullying detection, few studies have focused on analyzing texts written in languages other than English. Among the studies that have, Ozel et al. [18] used the Turkish language in their investigation of cyberbullying detection. To generate an evaluation dataset for their experimental work, they collected streaming data from Twitter. Each tweet was given its own vector using the bagof-words approach and classified using a variety of machine learning techniques (support vector machine, na¨ıve Bayes, C4.5 and KNN) to determine whether the posts involved mistreatment. In terms of F-measure, the Naive Bayes classifier significantly outperformed other classification techniques, with an accuracy rate of 79%.

Salawu et al. [19] presented a systematic review on cyberbullying detection approaches. They divided the existing approaches into four categories based on their substantial literature review: supervised learning, lexicon-based, rule-based, and mixedinitiative approaches. Supervised learning-based techniques commonly use classifiers such as SVM and naive Bayes to create predictive models for cyberbullying detection. Lexiconbased techniques identify cyberbullying using word lists and the presence of words within the lists. Mixed-initiative approaches combine human-based reasoning with one or more of the above-mentioned approaches to identify bullying. Rulebased approaches compare text to predetermined rules to identify bullying. The authors discovered two significant obstacles in cyberbullying detection research: the shortage of labeled datasets and academics' failure to take a holistic approach to cyberbully while creating detection systems. Their study effectively presents the current state of cyberbullying detection research with traditional ML techniques.

Rosa et al. [20] analyzed the existing research on automatic cyberbullying detection in depth. Their findings revealed that cyberbullying is frequently misinterpreted in the literature, resulting in erroneous systems with limited real-world utility. Furthermore, there is no standard methodology for evaluating these systems, and the natural imbalance of datasets continues to be an issue. They identified the future trend of research on the issue toward a position more consistent with the phenomenon's description and depiction, allowing future systems to be more practical and focused. Al-Garadi et al. [23]

studied existing publications to detect aggressive behavior using ML approaches. They summarized and recognized the critical factors for detecting cyberbullying through ML techniques, especially supervised learning. For this purpose, they have utilized accuracy, precision-recall and f-measure to determine the area under the curve function for modeling the behaviors in cyberbullying.

Elsafoury et al. [22] reveal some challenges and constraints of cyberbullying detection. Their paper represents a systematic literature review on automated cyberbullying detection that wraps all the steps in the ML pipeline. They also demonstrate that utilizing slang-based word embedding improves the detection of cyberbullying.

Kim et al. [21] give a thorough analysis of the past ten years of computational research concentrating on developing ML models for cyberbullying detection. A saturated corpus of 56 papers examined how humans are involved and considered directly or indirectly in building these detection algorithms. The authors focused on current algorithms' congruence with theories of cyberbullying. They then examined if and how current algorithms have incorporated humans. Finally, they shed insight into how academics have envisioned using current detection algorithms. Their evaluation reveals essential gaps in this research area due to the lack of human-centeredness in algorithm creation.

A comparison of automated cyberbullying detection methods, including data annotation, preprocessing, and feature engineering, is presented in the study by Al-Harigy et al. [24]. Emoji use in cyberbullying detection and the application of self-supervised learning to annotation are also covered. Due to the detrimental effects of cyberbullying, particularly on social media where anonymity can foster hate speech and cyberbullying, the paper emphasizes the need for efficient cyberbullying detection.

Nahar et al. [25] also experimented with clustering messages as part of the detection process. They used Kernelbased Fuzzy C-Means (K-FCM) to cluster the data by evaluating the features of a post and their relevance to a document class with the aim of identifying natural groupings. A Fuzzy SVM model was then used to classify each post using the membership matrix generated by K-FCM. This design was aimed at eliminating the inherent noise in social media data, thus improving the accuracy of the detection process. In another experiment, they adopted a semi-supervised learning approach that supplemented an initial training sample with additional training data extracted from unlabelled data. A linear compression voting function was then used to combine the outputs of Natıve Bayes and Stochastic Gradient Descent classifiers to decide if a post is bullying or not and to enlarge the training set with the labelled output from the classifiers. Like Nahar et al. [25], Sood et al. [26], [27], and Mangaonkar et al. [28] also introduced voting functions to determine the optimal configuration for cyberbullying detection. Sood et al. [26], [27] developed three profanity detection systems based on three separate features, namely a profanity dictionary, Levenshtein Edit Distance, and Bag-of-words. The profanity dictionary was based on a user-compiled list on phorum.com7 and noswearing.com. The second system used this profanity list in addition to an edit distance calculator to correct for misspellings. To eliminate false positives, the system checks the words against an English dictionary and a list of names. For example, an edit distance calculator will match 'shirt' to the profane term 'shit' and flag 'shirt' as an offensive term but, by consulting the dictionary, the system will identify the word 'shirt' as not being profanity. The third detection system is an SVM classifier that uses bigrams and word stems as features. Running a series of experiments using the three detection systems in various permutations, they obtained their best overall results using a configuration that combines the output of all three systems in an "OR" operation—i.e., if a comment is flagged as profanity by any of the three systems—and the most precise combination used the SVM-based system "AND" either the profanity list or the Levenshtein distancebased system.

For their cyberbullying detection system, Zhao and Mao [29] experimented with an SDA (Stacked Denoising Autoencoders) [30] variant called Semantic-enhanced Marginalized Stacked Denoising Autoencoders (smSDA) and what they termed Embedding enhanced Bag-of-Words (EBoW) [29]. They created an initial list of insulting words and used word embeddings to retrieve, from the corpus, words that are most similar to the insulting words. Their approach allowed a Linear SVM classifier to learn additional textual features that would otherwise have been deemed of little relevance. For example, the term "paki" in the phrase "be a good paki and say hello" is an ethnic slur but one that may not be selected as a feature if it is sparsely used within the corpus; if, however, "paki" co-occurs with other known cyberbullying words somewhere else within the corpus – for example, in a phrase such as "you are nothing but a f**king paki"—then this co-occurrence with a known profane word (i.e., "f**king") is used to promote "paki" to relevance as a feature. A system such as this can benefit from Parime and Suri's [31] proposal for a dynamically-sourced profane wordlist that is regularly updated from online resources to ensure that new offensive words are captured as they are coined.

Hosseinmardi et al. [32] proposed a cloud-based architecture for a scalable detection system for a large social network platform like Instagram. They used n-grams as input features to an SVM classifier and network-based features such as "number of followers", "number of followings", and "number of likes" alongside image features to a Natıve Bayes classifier, and found the Natıve Bayes classifier to be four times faster in predicting cyberbullying instances that the SVM. Rafiq et al. [33] also used a Natıve Bayes classifier along with AdaBoost, Decision Tree, and RandomForest classifiers to detect cyberbullying instances in Vine; they achieved a 76.39 percent accuracy with AdaBoost using unigrams, comments, profile and media information as features.

Nahar et al. [34] included sentiment features generated by applying Probabilistic Latent Semantic Analysis (PLSA) [35] to bullying posts alongside BoW features to train a Linear SVM classifier. They found cyberbullying detection improved with the inclusion of sentiment features compared to when only BoW features were used. Nahar et al. [36] achieved even better results by substituting a weighted TFIDF scheme for the bag-of-words (BoW) feature and used Latent Dirichlet Allocation (LDA) [37] instead of PLSA to identify sentiment features. Sanchez and Kumar [38] used a Natıve Bayes classifier on tweets extracted by querying Twitter for homophobic slurs and then detected tweets with negative polarity. While such techniques have been successfully used to detect cyberbullying instances, they are rarely sufficient on their own to accurately and consistently identify bullying episodes.

Munezero et al. [39] theorised that including sentimentbased features would improve the detection of anti-social documents. Thus, they expanded on their earlier work [40] by introducing emotion-based features to three classifiers, namely Natıve Bayes, SVM, and J48 classifiers. The effect of the inclusion of these features was, however, marginal compared to earlier experiments performed without sentimentbased features [40]. This inability of isolated sentiment analysis techniques to accurately detect cyberbullying can be inferred from the work of Xu et al. [41]. They trained four text classifiers (Natıve Bayes, SVM (linear), SVM (RBF) and Logistic Regression) on a Twitter corpus to identify bullying tweets and the roles played by people referenced within the tweets. By reviewing a subset of the extracted tweets, they detected seven emotions in the tweets, namely anger, embarrassment, empathy, fear, pride, relief, and sadness, and found that, while fear is the emotion most expressed in the tweets [41], it is often jokingly expressed. It would appear from our review that, when used in isolation for cyberbullying detection, sentiment analysis techniques struggle to distinguish between genuine emotions and those sarcastically expressed in bullying messages. We found that mixed-initiative approaches (discussed in a later section) provide a way to improve sentiment-based (and other) cyberbully detection approaches by injecting human-based logic into the detection process.

Following on from their 2011 work, Dinakar et al. [42] attempted to detect indirect bullying messages by incorporating common sense reasoning into their detection system. The common sense reasoning was implemented as a set of over 200 assertions converted into a sparse matrix representation of concepts versus relations (referred to as their BullySpace Knowledgebase). For each document in the dataset, a set of concepts were extracted and compared to the canonical concepts represented in the BullySpace Knowledgebase. Thus, a message such as “did you go lipstick shopping with your mum today” sent to a heterosexual male will be matched to the assertion “lipstick is used by girls” and then flagged as an instance of implicit cyberbullying indicative of homophobic sentiments. This method is an example of a mixed-initiative approach to cyberbullying detection, allowing the inclusion of human-based reasoning within the detection process. A bullying message such as this will normally go undetected in many traditional cyberbullying detection systems as it contains neither profanity nor negative sentiments. 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 realworld human knowledge.

Mancilla-Caceres et al. [43], [44] also studied user interactions within a virtual environment. They created a social computer game that required players to create teams and work collaboratively together to perform tasks. Using 5th grade students as case studies, they observed the students’ behaviours within the game and compared this to the results of a survey administered by cyberbullying experts to the same group of students prior to the game. By analysing interactions within the game, they discovered a collective attempt by a number of students to bully another student. Interestingly, none of the bullies were flagged by the cyberbullying experts as exhibiting bullying tendencies from the analysis of the survey responses. While such interactions within games and virtual worlds as studied by MancillaCaceres et al. [43], [44] and [44] offer an interesting insight into cyberbullying behaviour, care should, however, be taken when interpreting such data because certain seemingly inappropriate behaviour may be normal within a game-playing context. For example, within multi-player gaming worlds such as Call of Duty and World of Warcraft, players will often ridicule opposing players (referred to as “trash talk”) in an attempt to force an error.

Sabina T, Lise G, Yunfei C, and Yi Z [46] proposed A Socio-linguistic Model for Cyberbullying Detection propose a socio-linguistic model which jointly detects cyberbullying content in messages, discovers latent text categories, identifies participant roles and exploits social interactions. While method makes use of content that is labeled as bullying, it does not require category, role or relationship labels. Furthermore, as bullying labels are often subjective, noisy and inconsistent, an important contribution of this paper is effective methods for leveraging inconsistent labels. Rather than discard inconsistent labels, evaluate different methods for learning from them, demonstrating that incorporating uncertainty allows for better generalization. The proposed socio-linguistic model achieves an 18% improvement over state-of-the-art methods.System develop a series of probabilistic models of increasing sophistication. It build these models with Probabilistic Soft Logic (PSL), a recently introduced highly scalable probabilistic modeling framework.First model makes use of text, sentiment and collective reasoning. Next, incorporate seed-words and latent representations of text categories. Finally, make use of social information by inferring relational ties and social roles. This models are evaluated on a dataset of youth interactions on the social media platform Twitter. Twitter has emerged as a fertile environment for bullying. Twitter’s ability to provide a veil of anonymity can facilitate bullying. Whittaker and Kowalski found that though fewer survey participants used Twitter (69.4%) compared to Facebook (86.5%), a higher percentage of participants experienced cyberbullying on Twitter (45.5%) than Facebook (38.6%) (and other platforms). System compare this model to a baseline N-Grams model. This model is comparable to standard bag-of- ngrams approaches. Additionally, compare to an implementation of a state-oftheart approach. The contributions include strategies for learning from uncertain annotations and linguistic models which demonstrate the utility of domain knowledge and collective reasoning.

Finally,M Ramya and J Alwin Pinakas[47] studied different type of feature selection for text classification.Text categorization is the task of deciding whether a document belongs to a set of pre specified classes of documents. Automatic classification schemes can greatly facilitate the process of categorization. Categorization of documents is challenging, as the number of discriminating words can be very large. Many existing algorithms simply would not work with these many numbers of features. For most text categorization tasks, there are many irrelevant and many relevant features. The main objective is to propose a text classification based on the features selection and preprocessing thereby reducing the dimensionality of the Feature vector and increase the classification accuracy. In the proposed method, machine learning methods for text classification is used to apply some text preprocessing methods in different dataset, and then to extract feature vectors for each new document by using various feature weighting methods for enhancing the text classification accuracy. Further training the classifier by Naive Bayesian (NB) and K-nearest neighbor (KNN) algorithms, the predication can be made according to the category distribution among this k nearest neighbors. Experimental results show that the methods are favorable in terms of their effectiveness and efficiency when compared with other. Keywords– Feature selection, K-Nearest Neighbor, Na¤ıve Bayesian, Text classification.Automated text classification is a particularly challenging task in modern data analysis, both from an empirical and from a theoretical perspective.Feature selection, i.e., selecting a subset of the features available for describing the data before applying a learning algorithm, is a common technique for addressing this last challenge. It has been widely observed that feature selection can be a powerful tool for simplifying or speeding up computations, and when employed appropriately it can lead to little loss in classification quality. Nevertheless, general theoretical performance guarantees are modest and it is often difficult to claim more than a vague intuitive understanding of why a particular feature selection algorithm performs well when it does. Indeed, selecting an optimal set of features is in general difficult, both theoretically and empirically; hardness results are known, and in practice greedy heuristics are often employed. One recommendation to mitigate bias is explicitly preparing annotators. This leads to another difficulty, namely the availability (or lack thereof) of reliably annotated data. A factor that contributes to this proble is that there is no universally accepted definition of hate speech (a statement many publications would agree on, let alone one that is productive. One can point at a United Nations report for definition , would however argue that it does not satisfy the criteria of being a universally accepted productive definition on several accounts. For one, the recommendations in said document are not legally binding, thus their implementation in all member countries is not.

III. CONCLUSION

The prevalence of cyberbullying on online platforms and social media necessitates effective detection systems. The developed deep learning classifiers, particularly the hybrid deep learning CNN-BiLSTM and single BiLSTM classifiers, have demonstrated promising performance in identifying and classifying instances of cyberbullying. The accuracy of the hybrid deep learning CNN-BiLSTM and single BiLSTM classifiers was evaluated using binary and multiclass classification datasets. The results showed that the BiLSTM classifier outperformed the CNN-BiLSTM classifier in detecting aggressive or non-aggressive bullying. However, it is important to note the limitations of this study, including the limited scope of English-language datasets and the issue of overfitting with the binary class dataset. Future research should focus on developing state-of-the-art transformer models for online cyberbullying detection using multilingual datasets. The proposed cyberbullying detection system prevents individuals from becoming victims of cyberbullying and contributes to a safer online environment. Further enhancements and updates are necessary to keep up with the evolving nature of cyberbullying. In conclusion, the proposed cyberbullying detection system using deep learning approaches shows promise in accurately identifying instances of cyberbullying on online platforms. Future research should focus on addressing the identified gaps and challenges to enhance the effectiveness and applicability of such systems in combating cyberbullying.

ACKNOWLEDGMENT

We wish to avail this opportunity to acknowledge our profound debt and extend our sense of guidance to our Guide, Prof. Ms Vidya hari for his valuable guidance, advice, and encouragement that he led to the successful completion of this project. We also take this opportunity to express our profound gratitude and deep regards to our project coordinator Mr. Ajmal E B, for all his effort, time and patience in helping us to complete the project successfully with all his suggestions and ideas. Our sincere thanks to our honorable Principal, Dr. Amar Nishad T. M for his genuine support and for providing us with the necessary facilities to carry out the project work. We are exceedingly grateful to our Head of the Department, Dr.Rekha lakshmanan Kumar for his co-operation in this project's completion. We are also thankful to our parents for providing us with financial assistance and moral support for project completion.

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma, Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760, 2017.

[2] Md Abul Bashar and Richi Nayak, Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019), 2019.

[3] Pete Burnap and Matthew L Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy Internet, 7(2):223–242, 2015.

[4] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi, Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), pages 86–95, 2017.

[5] Shimaa M Abd El-Salam, Mohamed M Ezz, Somaya Hashem, Wafaa Elakel, Rabab Salama, Hesham ElMakhzangy, and Mahmoud ElHefnawi, Performance of machine learning approaches on prediction of esophageal varices for egyptian chronic hepatitis c patients. Informatics in Medicine Unlocked, 17:100267, 2019.

[6] Paula Fortuna and S´ergio Nunes, A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4):1–30, 2018.

[7] Purnama Sari Br Ginting, Budhi Irawan, and Casi Setianingsih, Hate speech detection on twitter using multinomial logistic regression classification method. In 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), pages 105– 111. IEEE, 2019.

[8] Yin, D.; Xue, Z.; Hong, L.; Davison, B.D.; Kontostathis, A.; Edwards, L, Detection of harassment on web 2.0. In Proceedings of the Content Analysis in the WEB, Madrid, Spain, 21 April 2009; pp. 1–7.

[9] Reynolds, K.; Kontostathis, A.; Edwards, Using machine learning to detect cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, Washington, DC, USA, 18–21 December 2011; pp. 241–244.

[10] Modha, S.; Majumder, P.; Mandl, T.; Mandalia, Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. Expert Syst. Appl. 2020, 161, 113725.

[11] Dinakar, K.; Reichart, R.; Lieberman, Modeling the detection of textual cyberbullying. In Proceedings of Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 6–9 June 2022.

[12] Dadvar, M.; Jong, F.D.; Ordelman, R.; Trieschnigg, Improved cyberbullying detection using gender information. In Proceedings of the Title of host publicationProceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, 24 February 2012.

[13] Kontostathis, A.; Reynolds, K.; Garron, A.; Edwards, Detecting cyberbullying: Query terms and techniques. In Proceedings of the 5th Annual Acm, Web Science Conference, online, 2 May 2013; pp. 195–204.

[14] Ptaszynski, M.; Masui, F.; Kimura, Y.; Rzepka, R.; Araki, Extracting patterns of harmful expressions for cyberbullying detection. In Proceedings of the 7th Language Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'15), The First Workshop on Processing Emotions, Decisions and Opinions, Pozna ´n, Poland, 27–29 November 2015; pp. 370–375.

[15] Zhang, X.; Tong, J.; Vishwamitra, N.; Whittaker, E.; Mazer, J.P.; Kowalski, R.; Hu, H.; Luo, F.; Macbeth, J.; Dillon, Cyberbullying detection with a pronunciation based convolutional neural network. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 740–745.

[16] Chavan, V.S.; Shylaja, Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 10–13 August 2015; pp. 2354–2358.

[17] Squicciarini, A.; Rajtmajer, S.; Liu, Y.; Griffin, Identification and characterization of cyberbullying dynamics in an online social network. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 25–28 August 2015; pp. 280–285.

[18] Ozel, S.A.; Sarac¸, E.; Akdemir, S.; Aksu, Detection of cyberbullying on social media messages in turkish. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 366–370.

[19] Salawu, S.; He, Y.; Lumsden, Approaches to automated detection of cyberbullying: A survey. IEEE Trans. Affect. Comput. 2017, 11, 3–24. [CrossRef]

[20] Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simao, A.V.; Trancoso˜ , Automatic cyberbullying detection: A systematic review. Comput. Hum. Behav. 2019, 93, 333–345. [CrossRef]

[21] Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P.J.; De Choudhury, A Human-Centered Systematic Literature Review of Cyberbullying
Detection Algorithms. Proc. ACM Hum.-Comput. Interact. 2021, 5, 1–34. [CrossRef]

[22] Elsafoury, F.; Katsigiannis, S.; Pervez, Z.; Ramzan, When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. IEEE Access 2021, 9, 103541–103563.

[23] Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani,
Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. IEEE Access 2019, 7, 70701–70718.

[24] Al-Harigy, L.M.; Al-Nuaim, H.A.; Moradpoor, N.; Tan, Building toward Automated Cyberbullying Detection: A

Comparative Analysis. Comput. Intell. Neurosci. 2022, 2022, 4794227.

[25] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, “Semi-supervised learning for cyberbullying detection in social networks,” Databases Theory App

[26] S. O. Sood, J. Antin, and E. F. Churchill, “Using crowdsourcing to improve profanity detection,” in Proc. AAAI Spring Symp. Wisdom Crowd. Stanford, 2012, pp. 69–74.

[27] S. O. Sood, E. F. Churchill, and J. Antin, “Automatic identification of personal insults on social news sites,” J. Amer. Soc. Inf. Sci. Technol., vol. 63, no. 2, pp. 270–285, 2012b.

[28] A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in Twitter data,” in Proc. IEEE Int. Conf. Electro/Inf. Technol., 2015, pp. 611–616.

[29] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” in Proc. 17th Int. Conf. Distrib. Comput. Netw., 2016, Art. no. 43.

[30] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” J. Mach. Learn. Res., pp. 3371–3408, Dec. 2010.

[31] S. Parime and V. Suri, “Cyberbullying detection and prevention: Data mining and psychological perspective,” in Proc. Int. Conf. Circuit Power Comput. Technol., 2014 , pp. 1541–1547.

[32] H. Hosseinmardi, S. A. Mattson, R. Rafiq, R. Han, Q. Lv, and S. Mishra, “Poster: Detection of cyberbullying in a mobile social network: Systems issues,” in Proc. 13th Annu. Int. Conf. Mobile Syst. Appl. Serv., May 18–22, 2015, pp. 481–481.

[33] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, “Careful what you share in six seconds: detecting cyberbullying instances in Vine,” in IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, 2015, pp. 617–622.

[34] V. Nahar, S. Unankard, X. Li, and C. Pang, “Sentiment analysis for effective detection of Cyber Bullying,” in Proc. Asia-Pacific Web Conf., 2012, pp. 767–774.

[35] T. Hofmann, “Probabilistic latent semantic analysis,” in Proc. 15th Conf. Uncertainty Artif. Intell., Jul. 30–Sep. 1, 1999, pp. 289–296.

[36] V. Nahar, X. Li, and C. Pang, “An effective approach for cyberbullying detection,” Commun. Inf. Sci. Manage. Eng., vol. 3, no. 5, 2013, Art. no. 238.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[38] H. Sanchez and S. Kumar, “Twitter bullying detection,” Int. J. Eng. Res. Appl., vol. 12, pp. 15–22, 2011.

[39] M. Munezero, C. S. Montero, T. Kakkonen, E. Sutinen, M. Mozgovoy, and V. Klyuev, “Automatic detection of antisocial behaviour in texts,” Informatica. Special Issue: Adv. Semantic Inf. Retrieval, vol. 38, no. 1, pp. 3–10, 2014.

[40] M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev, and E. Sutinen, “Antisocial behavior corpus for harmful language detection,” in Proc. Federated Conf. Comput. Sci. Inf. Syst., Sep. 8-11, 2013 , pp. 261–265.

[41] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, “Learning from bullying traces in social media,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Language Technol., 2012, pp. 656–666.

[42] K. Dinakar, et al., “You too?! mixed-initiative LDA story matching to help teens in distress,” in Proc. 6th Int. AAAI Conf. Weblogs Soc. Media , June 4 – 7, 2012, pp. 74–81.

[43] J. Mancilla-Caceres, D. Espelage, and E. Amir, “A computer gamebased method for studying bullying and cyberbullying,” J. School Violence, vol. 14, no. 1, pp. 66–86, 2015.

[44] J. Mancilla-Caceres, W. Pu, E. Amir, and D. Espelage, “A computerin-the-loop approach for detecting bullies in the classroom,” Social Comput. Behavioral-Cultural Model. Prediction, vol. 7227, pp. 139–146, 2012.

[45] T. Bosse and S. Stam, “A normative agent system to prevent cyberbullying,” IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., Lyon, France, pp. 425–430, Aug. 2011.

[46] Sabina T, Lise G, Yunfei C, Yi Z, ”A Socio-linguistic Model for Cyberbullying Detection”, in the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),
2018.

[47] M Ramya and J Alwin Pinakas, Different type of feature selection for text classification. International Journal of Computer Trends and Technology, 10(2):102–107, 2014.