2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579625

CYBER-BULLYING DETECTION VIA TEXT MINING AND MACHINE LEARNING

Chetna Sharma
Electronics Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
chetna.sharma@spit.ac.in

Rahul Ramakrishnan
Electronics Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
rahul.ramakrishnan@spit.ac.in

Ayusha Pendse
Electronics Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
ayusha.pendse@spit.ac.in

Priya Chimurkar
Electronics Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
priya.chimurkar@spit.ac.in

Kiran T. Talele
Electronics Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
kiran.talele@spit.ac.in

***Abstract*—Cyber Bullying is one of the most recent evils of social media. With a boom in the usage of social media, the freedom of expression is being exploited. Statistics show that overall 36.5 percent people think they have been cyberbullied in their lifetime. These numbers are more than double of what they were in 2007, and there is an increase from 2018-19, suggesting we are heading in the wrong direction. Solutions to curtail this issue to a certain extent have already been deployed in the market. However, they possess limitations of usage, or simply do not use efficient algorithms. This paper aims at identifying cyberbullying at is origin, meaning when it is being drafted in real-time. Identifying traces of cyberbullying before the content is uploaded on the internet can help reduce circulation of hurtful messages. Using Machine Learning with the support of Natural Language Processing(NLP, results in better performance of cyberbullying detection.**

***Index Terms*—Cyberbullying, Machine Learning, Natural Language Processing, Text Mining, Twitter**

I. Introduction

The internet is the world's biggest platform for communicating, sharing ideas, and content delivery. Social media comes in various forms and websites to make this process interactive. Twitter, YouTube, Instagram, LinkedIn are few of the largest of these platforms.

The usage of these platforms boomed post year 2000. Various subjects discussed and various ideas shared are supported and criticised. However, this criticism has gone to the extent of bullying. Individuals sitting behind screens target innocents, which causes damage of reputation and image at the very least. Sections of people are targeted based on gender, orientation, caste, and even color of the skin.

According to studies, over half the students who identify as LGBTQ have experienced cyberbullying at some point [8]. Girls are more likely to become a victim of cyberbullying as compared to boys. Overall, 36% of girls have reported being cyberbullied as opposed to 26% boys [9]. The amount of cyberbullying that now takes place has caused health issues for those targeted. 64% of cyberbullying victims say it has affected their ability to learn and do not feel safe at school [10]. Victims (students) are more likely to have social, mental and behavioral problems at school [11].

Various solutions in the form of third-party applications or plug-ins have been deployed in the market. The problem with these tools is that they are based on a simple keyword matching technique and are not always accurate. Using lexicons is a common way to use databases for detection of abusive words. However, it limits the scope of the application and disregards the statements which may not use abusive words but have hurtful meanings.

This paper aims at identifying cyberbullying at its origin, that is when it is being drafted in real-time. It proposes the use of Machine Learning to classify comments as cyberbullying if they have offensive words or meanings. Before the message or comment is sent to the required person, the author is notified of the content of the message requesting him/her to edit it and avoid hurtful/hateful phrases or words. Embedding this system in the functioning of the social media applications instead of making it a plug-in forces all users to follow these guidelines. It can encourage users and can spread awareness about the after-effects of cyberbullying.

II. Literature Survey

Cynthia Van Hee, et al. [1], developed model using tokenization, PoS-tagging and lemmatization for preprocessing, word n-grams and lexicon features extraction. Models were developed for English and Dutch to test for language conversion and subsequent accuracy. Mohammed Ali Al-Garadi, et al. [2], gave a comparative analysis of using SVM, K clustering, Random forest and Decision Trees. They concluded that SVM worked best amongst the four machine learning models. Kshitiz Sahay,et al. [3], used data set obtained from Wikipedia, YouTube, Twitter. Count Vectors and TF-IDF vectors were created by defining n gram of up to 5 levels.The

ML algorithms used were Logistic Regression, SVM, Random Forest and Gradient Boosting. Homa Hosseinmardi, et al. [4], developed a system for deciding posts based on shortlisting words of caption. The paper suggested using image processing on Instagram posts for deciding emotional response or test response in case of text pictures.

Maral Dadvar, et al. [5], performed comparison of 4 models of DNN - CNN, LSTM, BLSTM and BLSTM with attention. SSWE and GloVe was used for word embedding. The models were implemented on 3 datasets and the data was oversampled with negative comments for higher accuracy. John Hani, et al. [6], performed classification using SVM and neural networks. B.Sri Nandhinia, et al. [7], classified data using Fuzzy logic and Genetic Algorithm. Michele Di Capua, et al. [8], followed an Unsupervised approach with Syntactic, Semantic, Sentiment analysis. Preprocessing as stop word removal, punctuation removal was done to generate word clusters. Social features were extracted. Convolutional neural network was applied using Kohonen map (or GHSOM). Noviantho, et al. [9], suggested using Text Mining for shortlisting messages.

Prabhu Trisha N [10], performed sentiment detection, context analysis, text matching to detect hurtful labels in images/videos. Li Bohan, et al. [11], used bidirectional recurrent neural network. Word segmentation processing was used for filtering to identify the 'attention value' of text. Zhang Paurui [12], developed a system for blocking information from reaching users.Feedback was requested from user to improve detection. Dhruv Ghulati, et al. [13], reviewed sentiments in relation to the unlabelled content by determining a similarity or a probability score for abusive qualities. Danyluk Nicholas G, et al. [14], used previously flagged data for causing a negative reaction. This text is compared with any new text, and flagged subset of words is omitted from the new text. This reaction of user is captured by camera when data is displayed. Alexander James H., et al. [15], used feedback mechanism with input analysis using a context analyzer which is then categorized into safe or unsafe behaviour. Etter, David Lee, et al. [16], developed a system and device for classifying content data by calculating an alert score wherein the score corresponds to offensive content detected.

Anders, Kelley, et al. [17], Devised a system that consists of a network between multiple servers and clients for collecting, analyzing and storing data. Social media content in a class undergoes rigorous topic, tone, velocity and latent class analysis. Toxicity is classified and defined into several sub branches of body shaming, cyber-bullying etc. Newstadt, Keith, et al. [18], devised a system that identifies the online interaction of the user, checks the emotional response using a computing device and performs the security action if the emotional response is outside the expected range. Day, II Rowland W., et al. [19], developed a system that intercepts communications via the communication module and then blocks or filters the intercepted content accordingly and sends an alert relating to the intercepted communications. Daniel Hodges, et al. [20], developed a method that monitors a triggering criterion and reports it to the user with at least one monitored communication having bullying content responsive to detecting the triggering criterion as indicated by the user. Schoebel, Todd, et al. [21], developed an application that provides multiple alternate response and reporting actions and the user can associate a contact list with each of the menu actions. It also sends a copy of hurtful messages to the primary contact list. Daniel Hodges, et al. [22], developed a social networking system that is queried corresponding to the user. The queried information is compared with predetermined criteria to determine a content category corresponding to the identifying information, and a report is provided including an indication of the determined content category.

A field survey conducted in the community is analysed in Fig. 1. It gathered responses of over 300 users, to give a brief of the public opinion about their views on cyberbullying and discrimination.As per the above references, multiple solutions have been developed in the past to address cyberbullying. However, majority of the solutions consisted of drawbacks such as limitation to ally engineered datasets which could become subjective to the creators instead of the users, brute-force matching of live data and database data that leaves a small margin for identifying errors that are minutely different from the database, or products that give a warning to the users after the harmful comment has been published. Inversely, the proposed solution in this paper is developed by comparing multiple machine learning models and selecting the best suited model for twitter database. Multiple preprocessing methods applied to the input data allow variance in usage of words while maintaining a similar meaning. In addition , the warning is given to the user before the comment is published, hence reducing the chances of ignorance and spreading awareness.

III. Research Methodology

The research conducted for the effects of cyber-bullying was based on a public poll. We conducted an online survey with over 350 respondents with the aim to understand the current notion of cyber-bullying among people and majorly covering all possible reasons of cyber-bullying or basis for targeting certain sections of the society. The survey covered age distribution, gender discrimination, racism, hostile activity, xenophobia, body shaming, religion or social status of a a person. It was concluded that nearly 61% users agreed on women being subject to cyber-bullying more than men. Body shaming and sexual orientation were the top concerns for cyber-bullying.

Twitter was selected as the ideal platform to extract raw data. Twitter is one of the leading platforms for discussing all kinds of societal issues, and hence gives a large amount of views of people on various topics.

This paper suggests solutions for cyberbullying detection using raw tweets from Twitter as a data set. Text mining is used to perform the tweet extraction in large numbers of at least 30 thousand tweets. Training and testing datasets were used for the purpose of evaluation of various Machine Learning models.

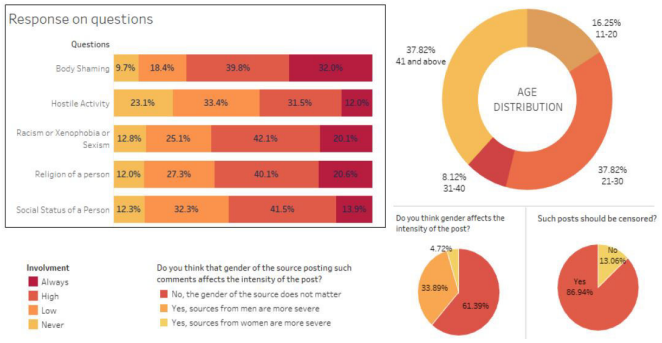

Fig. 1. Analysis of Field Survey conducted

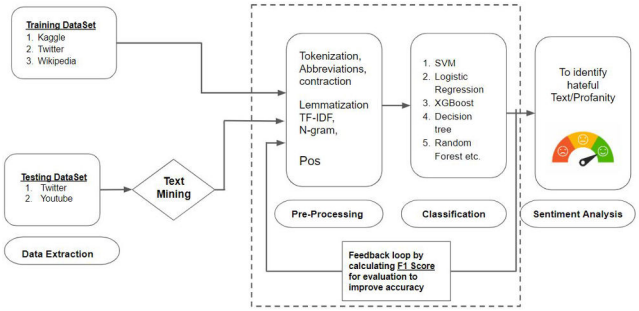

Fig. 2. Functional Block Diagram of method followed

Fig. 2. gives the Functional Block Diagram of the method followed. The dataset includes those extracted from Kaggle, Twitter, Wikipedia and YouTube. The datasets are divided into two types, those for training and testing. The training dataset was further divided into training and validating dataset in the ratio 3:2 which incorporated upto 30,000 tweets. The testing datasets needs to be extracted from the platforms via text mining for a real time usage of the system. Both the datasets pass through preprocessing techniques and various ML models. The F1 score and accuracy of the models is used as feedback to the system for improved performance. The conclusion drawn from the output is finding the most efficient ML model and getting its optimum efficiency.

IV. Pre-Processing Techniques

The raw tweets imported from Twitter were used as the testing dataset. Very few datasets which are primarily labelled as bullying or non-bullying were available. However, each of these datasets are not 100% accurate and have an accuracy of 95-96% themselves. This causes in a certain number of false positives and false negatives in the training dataset.Thus, these datasets were manually updated. These datasets required pre-processing before passing through multiple ML models. This preprocessing is necessary to clean the raw data specifically. When raw tweets of various users are imported, it is embedded with multiple system generated characters and encoding.

Characters are replaced with their hexadecimal values, whereas all the @" references made by users are not a part of the actual content to be classified. This calls for data cleaning for various conditions. These conditions were manually set, to remove hexadecimal values, system generated characters for new line or tab, and to remove @ references. It was concluded that numeric values add less value to the content when statement comprehension is required.

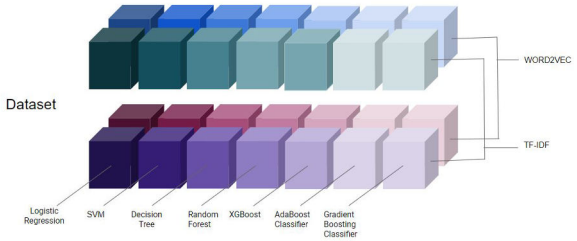

Fig. 3. Distribution of ML models with Datasets

More focus should be on the words used, and the meaning of the statements, instead of the numeric values which are majorly used for emphasis or statistics. Words shorter than 3 letter long were omitted. These words often include 'is, of, in, it, he, to etc.'. These set of words are used more for grammatical coherence in the statements and less to add sentimental value to the statement. hexadecimal character patterns and system encoding was removed but however hashtags were retained as they added more value to the statements and proved helpful to easily classify such statements. This was followed by 2 key processes - Vectorization and Lemmatization. Vectorization was performed prior to sending the training and testing dataset through the ML models - TF-IDF and Word2Vec. Lemmatization helps to break down various forms of a single word to its root meaning. This helps generalise the words and reduces complexity. After these steps of preprocessing, various machine learning models were studied and identify 7 ML models to compare for functionality. These models were chosen on the basis of popularity, ease of use, back end working and research results of various authors.

V. Machine Learning Models Used

Fig. 3. gives the list of ML models namely - Logistic Regression, SVM, Random Forest, XGBoost, Decision Trees, ADA Boost Classifier, Gradient Boosting Classifier.

VI. WebPage Development

Fig. 4. Shows the layout of the HTML page. The web page for the implementation of this project has been developed using HTML and CSS. The main directory containing all the files consists of a sub folder known as templates which contains all the HTML files. The entire web application as of now consists of 2 pages. The first template or file being the Review Form page which is where the user will enter their post or comment to be checked.

The page also consists of a submit button which when pressed will direct the user to the second template. This

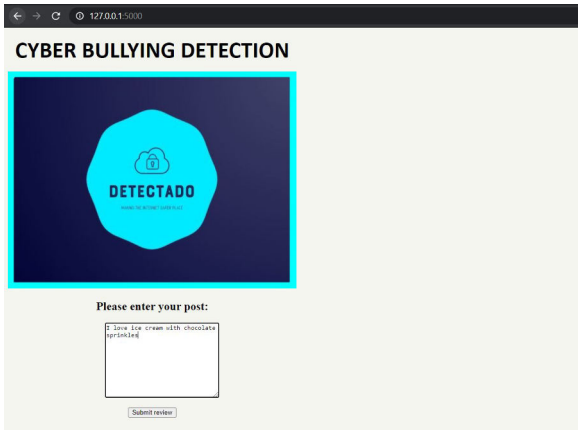

Fig. 4. HTML page to enter post

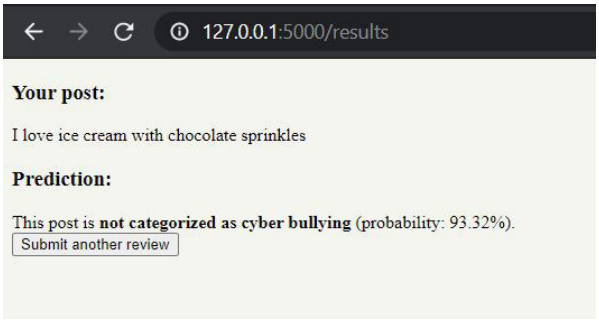

Fig. 5. HTML page to see result of post

page as given in Fig. 5. will display the post written and the prediction of whether this post would be categorized as cyberbullying or not and the probability of this prediction. The third template which actually is not visible on screen but is required for the functioning of these pages the standard code required for functioning of these two pages. This template assists Flask to link the ML model to the Web page, we write a macro that renders a field with label and a list of errors if there are any in our code.

To connect these HTML pages to the ML model we have used Flask framework. The Flask main directory consisted of a folder template for the HTML pages, a folder static for the CSS, a pkl objects folder and the flaskapp python file. The main python file i.e. flaskapp was run via the command prompt and ran on a local server 127.0.0.1/5000 which is given by the command prompt.

VII. Results and Analysis

The model is able to accurately classify 90% of all tweets passed through it. It positively classifies tweets regarding various topics such as racism, sexism, fat-shaming, politics, profanity, and hateful abusive language. A wide range of cyberbullying is covered to increase the universality of the model. As these topics are more saturated with cyberbullying all over the world, they were chosen. The model correctly classifies tweets falling under the above topics.

The seven Machine Learning models tested were:

- Logistic Regression
- SVM
- Random Forest
- XGBoost
- Decision Trees
- ADA Boost Classifier
- Gradient Boosting Classifier

F1 score of each of the models as well as accuracy were calculated for the datasets used. For understanding the usage of the ML models better, 2 pairs of datasets were used. In Table 1, a relatively larger dataset including 30,000 tweets.

TABLE I
ML Models Results - Dataset 1

Model Name	Vectorization	F1 Score
Logistic Regression	TF-IDF	0.606
Logistic Regression	Word2Vec	0.620
SVM	TF-IDF	0.605
SVM	Word2Vec	0.632
RANDOM FOREST	TF-IDF	0.5
RANDOM FOREST	Word2Vec	0.513
XGBoost	TF-IDF	0.75
XGBoost	Word2Vec	0.682
Decision Trees	TF-IDF	0.62
Decision Trees	Word2Vec	0.43
ADA Boost Classifier	TF-IDF	0.5
ADA Boost Classifier	Word2Vec	0.459
Gradient Boosting Classifier	TF-IDF	0.49
Gradient Boosting Classifier	Word2Vec	0.48

It was concluded that WORD2VEC gives better accuracy than TF-IDF. A larger Dataset, with a ratio of Training : Testing = 10:1 gives a more comprehensive understanding of ML model performance, with an increased scope for improvement. SVM is able to give a comparatively better output on a small data set. XGBoost gives the best output, when adjusted with parameters on the larger dataset but with larger computational time. Considering all different parameters such as accuracy score, precision score, F1 score and computational time and changing various parameters like regularization parameter and kernel SVM proved to give best results.

VIII. Limitations

The major limitations of this project revolves around acquiring labelled datasets for training of Machine Learning models. This project is fairly new and there is a scarcity of larger datasets which are accurately labelled. Most of the ones available are results of projects made using smaller datasets, which result in not very accurate results. With a dataset of even 15,000 tweets, an accuracy of 95% results into unreliable output. Finding datasets is expensive, especially ones which are 100% correct. This is because these datasets are labelled manually which require a lot of human effort. Progress is being made slowly and steadily for the above reason, which will yield into much better results in the near future.

IX. Future Scope

The project could have a number of applications in the future which are not only restricted to Detection of Cyber Bullying but also Email Spam Classifications, Fake News Detection ans so on and so forth. This classification can be further improved with a deeper focus on sentiment, semantic and syntactic analysis.Also the accuracy of such models is always improved the more it is trained. Secondly, creating a dynamic model which self updates the database for training will make this project much more self reliable and reduce the margin of error by minimizing human intervention

Another possible area of improvement could be to make the model work as a cyber-bully detector for all environments without having to train it with a new dataset everytime.

Another scope for improvement would be if the model also processed messages of vernacular language or of another language which was written or scripted in English. Use of emojis, emoticons, animojis and memojis could be included in this to improve the accuracy of this concept. On platforms such as Instagram the major method of communication is viz pictures which may have text embedded in them at times. Hence a method in which such category of posts or messages can be accounted for will make the project and model much more reliable, user friendly, effective, efficient and relevant.

X. Conclusion

The comparison of various ML models helps to conclude that XGBoost along with Word2Vec gives the best combination to classify various tweets into non-cyberbullying and cyberbullying. This model has an accuracy of 0.964 and F1 score of 0.847. This classification can be further improved with a deeper focus on sentiment, semantic and syntactic analysis. An integration with convolutional neural networks or deep learning will help increase the accuracy even further due to the complexity considered in the model. This model can be made an integral of various social media platforms to make it a necessary feature for all users. It can help reduce cyberbullying to a much greater extent and spread awareness about it.

It was also found that WORD2VEC gives better accuracy than TF-IDF. A larger Dataset, with a ratio of Training : Testing = 10:1 gives a more comprehensive understanding of ML model performance, with an increased scope for improvement. XGBoost gives the best output, when adjusted with parameters on the larger dataset but the computational time taken by xgboost is greater than svm. After the comparison of various ML models helps to conclude that SVM along with Word2Vec gives the best combination to classify various tweets into non-cyberbullying and cyberbullying. This model has an accuracy of 96.4%. This classification can be further improved with a deeper focus on sentiment, semantic and syntactic analysis. After testing with various different kernels available in SVM classifier, RGB kernel gives us the most accurate results in all the three parameters that is accuracy score, precision score and f1 score. An integration with convolutional neural networks or deep learning will help increase the accuracy even further due to the complexity considered in the model. This model can be made an integral of various social media platforms to make it a necessary feature for all users. It can help reduce cyberbullying to a much greater extent and spread awareness about it.

References

[1] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Veronique Hoste, 'Automatic Detection of Cyberbullying in Social Media Text', PLOS ONE Journal, 2018.

[2] Mohammed ali al-garadi, Nawsher khan , Ghulam murtaza, Ihsan ali, Hasan ali khattak, and Abdullah gani , 'Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms', IEEE Access,2019.

[3] Kshitiz Sahay, Harsimran Singh Khaira, Prince Kukreja, Nishchay Shukla, 'Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning', International Journal of Engineering Technology Science and Research, (IJETSR), 2018.

[4] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra,'Detection of Cyberbullying Incidents on the Instagram Social Network', University of Colorado Boulder, 2015.

[5] Maral Dadvar Kai Eckert,'Cyberbullying Detection in Social Networks Using Deep Learning Based Models, A Reproducibility Study', Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms', ECIR'18, 2018.

[6] John Hani, Mohamed Nashaat, Mostafa Ahmed, 'Social Media CyberBullying Detection Using Machine Learning', (IJACSA) International Journal of Advanced Computer Science and Applications, 2019.

[7] B.Sri Nandhinia , J.I.Sheebab, 'Online Social Network Bullying Detection Using Intelligence Techniques', International Conference on Advanced Computing Technologies and Applications (ICACTA- 2015), 2015.

[8] Michele Di Capua, Emanuel Di Nardo, Alfredo Petrosino, 'Unsupervised Cyber-Bullying Detection in Social Networks', 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016.

[9] Noviantho, Sani Muhamad Isa, Livia Ashianti, 'Cyberbullying Classification using Text Mining', 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 2017.

[10] Prabhu Trisha N, 'Method to stop Cyber-Bullying before it occurs', US9686217B2, December 17, 2015.

[11] Li bohan; Li xue, Wan Shuo, Wang Wenhuan, Wang Xueliang, Zhang Anman, 'Method and system for detecting cyberbullying', US20180295229A1, January 17, 2020.

[12] Zhang Pairui, 'anti-cyber bullying method', US20180295229A1, March 22, 2019.

[13] Dhruv Ghulati, Zeerak Waseem, 'Hate speech detection system for online media content', GB2572320A, October 02, 2019.

[14] Danyluk Nicholas G, Sehgal Kavita, Stamboni Diane M, Varghese Sneha m, Werner John S, Wu Sarah, 'COGNITIVE RECOGNITION AND FILTERING OF CYBERBULLYING MESSAGES', US20200028810A1, January 23, 2020.

[15] Alexander James H., Packer Alan, Rehfuss Paul Stephen, 'Detecting sexually predatory content in an electronic communication', US7444403B1, October 28, 2008.

[16] Etter, David Lee, France, Jason Otis, Ryan, Jr., James Ronald, 'System and method of detecting offensive content sent or received on a portable electronic device', US 20190266444A1, February 05, 2009.

[17] Anders, Kelley, Dunne, Jonathan, Fox, Jeremy R., Harpur, Liam S., 'SOCIAL MEDIA TOXICITY ANALYSIS', US 16/109696, February 27, 2020.

[18] Newstadt, Keith; Sokolov, Ilya, 'Systems and methods for analyzing emotional responses to online interactions', US 10419375 September 17, 2019.

[19] Day, II Rowland W., Roarke, Jacquez Partha, Sigler, Steven, Wise, Eric, 'DEVICES AND METHODS FOR IMPROVING WEB SAFETY AND DETERRENCE OF CYBERBULLYING', WO2015095597A8, June 25, 2015.

[20] Daniel Hodges, San Francisco, CA (US), Andrew Weiss, San Ramon, CA (US), Joseph Anakata, Alameda, CA (US), 'Communication monitoring system and method enabling designating a peer', US8788657B2, July 14, 2014.

[21] Schoebel, Todd, 'Cyber-Bullying Response System and Method', US 2014/0310191A1, October 16, 2014.

[22]Trisha N. Prabhu, Naperville, IL (US), 'METHOD TO STOP CYBER-BULLYING BEFORE IT OCCURS', US20150365366A1, December 17, 2015.

[23] Daniel Hodges, San Francisco, CA (US), Joseph Anakata, Alameda, CA (US), 'SYSTEM AND METHOD FOR IMPROVED DETECTION AND MONITORING OF ONLINE ACCOUNTS', US9571590B2, February 14, 2017.