A Comprehensive Study on Cyberbullying Detection Using Machine Learning Approach

Vaishali Malpe[1*], Shubhangi Vaikole[2]

[1]*Research Scholar, Computer Engineering Department, Datta Meghe College of Engineering, University of Mumbai, India*

[2]*Research Guide, Computer Engineering Department, Datta Meghe College of Engineering, University of Mumbai, India*

Abstract

Cyberbullying is an action where a person or a group of persons uses social networking sites on internet through smartphones, computers, and tablets to trouble, distress, hurt or harm another person. Cyberbullying occurs by sending, posting or sharing offensive or harmful texts, images or videos. It also involves activities such as sharing someone's personal or private information which causes feeling of awkwardness and shame, also humiliation. These actions are unlawful. With the increasing adverse impact of cyberbullying on society, it is necessary to find ways to detect this phenomenon. Automatically identifying bully words, emojis and audio/video features from online social platforms, especially micro-blogging site such as Twitter and video-sharing platform such as YouTube is an important research. This paper presents a collective and structured study to reconnoiter and assimilate research done in the field of detection of cyberbullying, also research gaps are illustrated in a legitimate manner. The study portrays a comprehensive systematic literature review of strategies proposed in the field of text-based and video-based cyberbullying. The survey relates to several machine learning methodologies and online social networking datasets used in previous studies and scope for improvement in detection of cyberbullying. This methodical analysis of the research work acts as an assistant for the researchers to discover the significant and compelling characteristics of cyberbullying detection techniques. Finally, issues and challenges in cyberbullying detection are highlighted and discussed.

Keywords: Online social media, cyberbullying, machine learning, natural language processing, deep learning

1. Introduction

Millions of youths are spending their time on social media devotedly and exchange information online. Social media has the ability to connect and share information with anyone at any time with many people simultaneously. There are more than 3 billion social media users globally. It is a web based platform which ever-changing and ever-evolving. Figure 1 depicts the statistics of popular social networking websites used worldwide as of January 2020, listed by number of functional users. According to National Crime Prevention Council (NCPC), cyberbullying is existed through internet where mobile phones, video game applications or other medium to send or post text, images or videos to hurt or embarrass another person intentionally. Cyberbullying can happen at any time throughout the day, in a week and outreach a person anywhere through internet. Cyberbullying texts, images or videos can be posted in an undisclosed way an

distributed instantly to a very wide audience. It can be difficult and sometimes impossible to track the sources of these posts. Deleting such messages after a period of time is also not possible. Many social platforms such as Twitter, Instagram, Facebook, YouTube, Snapchat, Skype and Wikipedia are the medium where cyberbullying is taking place. Some social media provide guide to prevent the cyberbullying. Facebook has the special section which describes how to report cyberbullying and block the user to prevent it. On Instagram if someone is sharing photos or videos that make user uncomfortable, user

can unfollow or block them. User can also report violations to Community Guidelines right from the app. Twitter recommends to block the user for his/ her inappropriate, abusive, offensive or threatening behavior. Cyberbullying is associated with social, emotional and academic problems and causing not only depression and detachment in teens, but also causing the threat of self-harm like suicidal behavior. Several organizations are working towards spreading the awareness about cyberbullying. Many researchers have lent their contributions in inventing the ways to early detect and prevent it. With the ability of machine learning algorithms to correctly classify data, and come up with accurate result, it has become the choice of many researchers to use them for detecting cyberbullying in association with natural language processing (NLP) techniques. Cyberbullying should be understood and treated from different perspectives. Automatic detection and prevention of such acts can help to tackle this problem. The effect of cyberbullying on multiple social media platforms cannot be neglected and hence requires serious attention to control these activities.
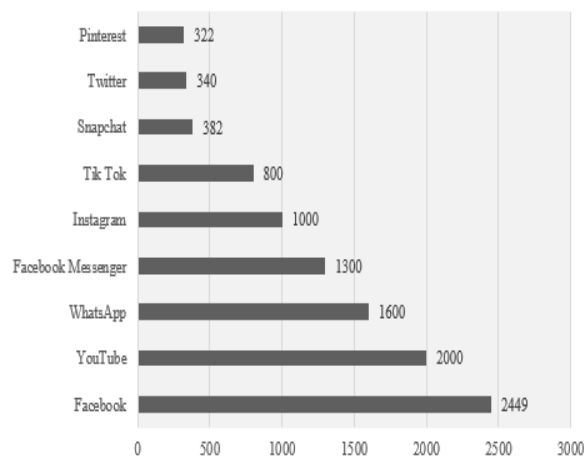
**Figure 1. Number of Functional Users on Various Social Networking Websites**

2. Background
2.1 Machine Learning

Machine learning is an application of artificial intelligence and refers to the ability to provide automatic learning and improve the results from experiences by identifying patterns. Machine learning uses existing algorithms and datasets and develop computer programs to provide adequate solution to the problem and use them to learn on their own. The process of learning starts with observations in data, recognizing the patterns in data and make advance decisions and use them in future based on the previously identified patterns. The prime aim is to make computers learn automatically without human involvement or aid and adjust results accordingly. Machine learning provides more accurate results in a faster way by analysing large amount of data. Machine learning is being used passionately and actively in industry today.

2.2 Cyberbullying

Cyberbullying is an act that exists where digital devices like smartphones, computers, and tablets are used. Cyberbullying can occur by viewing or participating in online forums on social media platforms or apps or gaming sessions or sharing negative, insulting or mean messages in the form of SMS or posts, or by sending or posting embarrassing images or videos. Such kind of posts are intended to humiliate the person. It can also include sharing one's personal or private information with others causing criminal behaviour. Figure 2 depicts different types of cyberbullying.
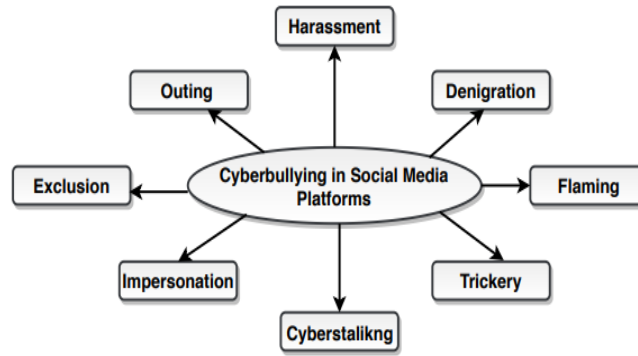
Figure 2. Various Ways of Cyberbullying on Social Media Platform

Many intricacies are involved to develop the system for such online circumstances. Following are the challenges identified [1].

1. Multifariousness of users with respect to their cultural norms and the study of population.
2. Fluctuating nature of problems as such phenomena are temperamental and they may start and stop unexpectedly.
3. The capacity of social media for the user to be anonymous and target the bullies easily without fearing consequences.
4. Different forms of bullying such as cyberaggression, sarcasm or trolling other than using abusive or profane language.

Following are the additional difficulties with respect to cyberbullying and cyberaggression prevention in Twitter [1].

1. As tweets are small texts with non-formal content having grammar and syntactic flaws, the natural language processing techniques for extracting text-based features and characterize interactions between users are not enough.
2. As every tweet has limited text, one tweet can be referred as a nonaggressive even if it contained aggressive context, whereas along with other tweets or multiple aggressive users' tweets it can be disregarded as a tweet.
3. Having spam or fake accounts can be considered as a bullying or aggressive action. Twitter is seen as a pool of numerous spam accounts.
4. Loss of privacy due to monitoring, forwarding to third parties (e.g., parents/admins), or removal of messages.
5. Willingness of victims to report cyberbullying incidents.
6. False reporting of cyberbullying instances.
7. Short text nature of social media posts.

3. Motivation

Cyberbullying is threatening and destructive act which may result in suicide attempts or negative impact can cause life-long harms to the victims. Above mentioned concerns urge the researchers to design the model to detect and prevent the abusive behavior on social media. The motivation to build an efficient multimodal detection system comes from the users' offensive behavior as it evolves in multiple social media platforms. Most of the existing work in this area have produced their results using supervised learning algorithms and researchers have focused on detecting cyberbullying based on textual information. This paper has organized the literature review of revolutionary research in the area of detection of text-based cyberbullying. Throughout the research it has been noticed that the

detection of audio and video cyberbullying is still a neglected area. The work sums up the existing research on the basis of extensive and systematic search available in literature.

4. Review of Literature

The study includes research work carried out by many researchers in the field of cyberbullying detection using machine learning approaches across multiple social media platforms. Most of the studies came out with text based cyberbullying detection and very few are based on video cyberbullying. Many studies have come up with the better accuracy with use of supervised learning algorithms, researchers have shown their interest in using Natural Language Processing techniques to improve the performance of the classifiers.

Despoina C. et al. in [1], the study has identified the challenges in detection of cyberbullying such as heterogeneity of users, transient nature of the problem, anonymity capability offered in social media, and multiple bullying forms beyond abusive language. The authors have considered the user, textual and network features to detect cyberbullying. Supervised machine learning algorithms have been used to classify the text as bully or non-bully.

In [2], text-based Convolution Neural Network (CNN) using fastText word embedding was built to identify toxic and abusive comments on social media platform and classifying them based on their toxicity level. It is concluded that fastText has provided more accurate results when dealing with slang, jargons, typing mistakes and short forms used in the posts. The model had outperformed when the datasize was large enough to split as training and testing data.

John M. et al. in [3] has used a supervised machine learning algorithms to detect and prevent cyberbullying. Several classifiers were used to train and identify the bullying actions. When the proposed approach on cyberbullying dataset was evaluated, it is showed that Neural Network performs better and achieved accuracy of 92.8 % and Support Vector Machine achieved 90.3 %. Also, Neural Network outperformed other classifies of similar work on the same dataset.

Lu Cheng et al. in [4] proposed XBully, a multimodal cyberbullying detection system. It was based on network representation learning. This system has taken into account the various feature types and those were handles by identifying representative mode hotspot. It was then mapped in a heterogeneous network. During the social session interaction different roles such as victims or bullies could be found out so that cyberbullying classification can be improved.

Cynthia et al. in [5] has built a classifier to detect indications of cyberbullying on social media platform which identifies different social roles involved in a cyberbullying interaction. Roles were discriminated in the annotation scheme which includes victim, bully, bystanders-defendant and bystanders-assistant. Linear support vector machine was used as a classifier. They demonstrated the method which can be used for languages easily. The experiments were performed in English and Dutch datasets.

The study presented in [6] evaluated both machine learning and deep learning based models and proved that deep learning based model has performed the better accuracy. At the same time these models needs huge amount of data to achieve the accuracy. Also the classification time in these model is more. The researchers also performed the classification of emoticons in different categories such abusive, sad, happy, etc. They have concluded that the image analysis, emojis and roles of victims and attackers would improve the cyberbullying classification. They have used the dataset from Twitter, Instagram and Ask.fm.

Sabina Tomkins et al. in [7] developed two models, domain-inspired linguistic model and a socio-linguistic model. The domain-inspired model has exploited the relationship between the word and the document. This is done to reduce the sparsity as the social media messages are usually short, misspelled and slang which are not suitable for generalization. If the dataset is limited, social-linguistic model is used which is used infer the relationship between participants to identify their roles. The parameters F-measure and recall has achieved better performance in socio-linguistic model as compared to latent linguistic model.

Hitesh S. et al. in [8] have used different supervised machine learning algorithms such as Logistic Regression (LR), Random Forest Classifier (RF), Support Vector Machine (SVM) and Gradient Boosting Machine. Comparison has been done to portray best performing classifier. Twitter and YouTube dataset were used which are publically available. The experiment was done feature stack which contained user, textual, network and lexical syntactic features. It is concluded that LR and RF Classifier achieved good results than Support Vector Machine and Gradient Boosting Machine.

In [9], both supervised and unsupervised approaches are used. The authors have identified the misspelled and censored words by using pronunciation of words. The Soundex algorithm was used to convert the words into their respective pronunciation code and feature vector was built. K-Means and Latent Dirichlet Allocation clustering techniques were used to detect cyberbullying whereas Naive Bayes Classifier and Support Vector Machine were used as classification algorithms. The work has been carried out on multiple datasets.

Sweta A. et al. in [10] have introduced the cyberbullying detection on different social media platform and developed a model that can be transferred from one platform to another to another to detect cyberbullying. The datasets used were Fromspring, Wikipedia and Twitter. They have developed four different Deep Neural Network models to solve the purpose. The models have used different features from literature along with data such as profile and social graph of the users to improve the performance of classifiers. Authors have observed that different words used for cyberbullying and their interpretation varies remarkably on different social networking sites.

Harsh Dani et al. in [11] calculated the sentiment score distribution of normal posts and bullying posts and compared these scores to verify the sentiment difference. This sentiment information and user-post relationships were used to build the framework which would detect the cyberbullying. The experiments were carried out on two real-world datasets MySpace and Twitter. Authors claimed that impact of sentiment information achieved more effective results in terms of accuracy.

Elaheh Raisi et al. in [12], have used participant-vocabulary consistency (PVC) which discovered which users are initiating the bullying and which users are victims. Simultaneously it has recommended the words that indicates bullying. Language-based and network based parameters are used to estimate how interaction on social networking platforms contains bullying. PVC has discovered instances of bullying and new bullying language. The language and participants used in social interactions was estimated by the model. PVC, on data sets, Twitter, Ask.fm and Instagram has proved its effectiveness in cyberbullying detection quantitatively and qualitatively.

In [13], Growing Hierarchical Self Organizing Map, an unsupervised approach has been used to build the model to detect cyberbullying on social media platform. The datasets Twitter, YouTube and Formspring were used. The work has been carried out based on semantic and syntactic communicational and social behavioural features. Model has

achieved admissible performance and authors claimed that it could be applied to develop the system where cyberbullying detection is important. Applications such parenting app or law agencies applications.

Vivek K. Singh et al. in [14], proposed a framework that has confidence score and social and textual features of posts that can be used to predict the roles of users as predators or victims. The roles of users involved in detection of cyberbullying. They have used Synthetic Minority Oversampling Technique on Twitter dataset. The proposed system was evaluated and compared with the systems built in literature which used same datasets and features. Authors claimed that their model has performed significantly well in detecting cyberbullying.

Rahat et al. [15] has built the model to discover video based cyberbullying. This is the first attempt made for detecting visual cyberbullying. Availability of literature for video based detection is very few. Rahat has considered the media session comprises of videos, its associated comments and expressions extracted from video. The labelling has been done through the survey conducted among experts who know the clear definition of cyberbullying. He has also stated the clear definition of cyberbullying and cyberaggression. Text based detection has also been implemented.

Similar procedure is followed in [16] using the dataset of images from Instagram and its associated likes and comments. Instagram is a mobile based image and video sharing application which allows user to like, comment and post text, images and videos and to follow other users. The proposed work had identified that many media session contains comments which represents cyber aggression but not cyberbullying. The authors had clearly mentioned that cyberbullying and cyberaggression are two separate concepts, however cyberbullying can be a part of cyberaggression.

The textual and social features are used in [17] to do classification. Social features represent the relationship between users and can identify sender and receiver. The comments posted with '@' were considered which signifies the relationship between sender and receiver as the symbol helps to address the user. Ego network and social network are used to derive the social features. It has been concluded that Social features are useful in detecting cyberbullying.

Cicero Nogueira et al. in [18] has performed the sentiment analysis of character-level, word-level and sentence-level representations. They presented the deep learning neural network model. The authors applied the approach to two different domains: the Stanford Sentiment Treebank (SSTb) and the Stanford Twitter Sentiment corpus (STS). For the STS corpus, a sentiment prediction accuracy of 86.4% was achieved. The main contributions of the paper were the idea of using convolutional neural networks to extract from character-to sentence-level features.

Vinita Nahar et al. in [19] employed the concept of link analysis to find the roles of active users such as predators and victims. The session-based one-class ensemble classifier was built to classify each entry as a 'bullying' or 'non-bullying'. The classifier has used small set of labelled data and huge unlabeled data. The experimental results showed that the ensemble classifier had outperformed the single window and fixed window approaches.

Maral Dadvar et al. in [20], authors used a combination of features which are based on contents of posts and features which are related to the user of account holder on social media platform. They showed that cyberbullying detection performance improved when more bullying-specific features were added. It improved the performance of the model

when the context information such as user's comment history and user's characteristics from his/ her profile were added. A bag of profane words, pronoun, profanity windows, and second person pronouns' frequency were the main contributing features. Incorporation of users' profile information improved the precision and the recall to 77% and 55% respectively.

Vinita Nahar et al. in [21], authors have proposed two feature selection methods, common features generated by bag-of-word approach, features extracted from both bullying and non-bullying messages and sentiment features generated by Probabilistic latent semantic analysis, features extracted from only bullying messages. Linear support vector machine was implemented to detect cyberbullying. Communication network of users was taken into account which includes predator and victims. Victim and predator are identified by calculating the score using HIT algorithm. Highest predator score represents a predator and highest victim score represents a victim. They have achieved very high accuracy.

Maral Dadvar et al. in [22] have adopted the gender-based approach to build the support vector machine classifier which would detect cyberbullying on social media platform. Authors believed that male use more profane and negative words compared to female. Therefore considering user's information such as age and gender along with contents of the post would improve the accuracy of cyberbullying detection. Each post on MySpace corpus was treated individually. They have targeted the posts which are shared by children and teens.

Study in [23] is followed by all the cyberbullying researchers. The researchers have used the Formspring data to carry out the work. Preprocessing has been done on data by removing repeating and unimportant word. The number of bad words and density of bad words are used as features. Supervised machine learning approach was used to do cyberbullying detection. The better accuracy was achieved for the small size data.

Following are the summarized findings which give new directions towards research.

1. Although many researchers have worked on cyberbullying detection on textual features, there is a room to consider more brilliant features or combination to enhance the detection. The available datasets do not provide any information about how severe the bullying is. If such information is made available, the cyberbullying detection models can be implemented to take different action depending on the seriousness of posts.

2. There is lack of study where profile of social media user is taken into account for feature extraction. This would be helpful to guess the user's behavior so as to achieve accurate classification.

3. In spite of being complex, Deep Neural Network has been proved as the technology which gives better performance. As sending images and videos is becoming popular among adolescents, image/video processing would be another important area for cyberbullying detection. Using Deep Neural Network, accurate video cyberbullying detection can be achieved which is still a neglected domain on social media platforms.

4. No study has come out with the audio classification where there is high possibility of using hate speech, offensive, profane words.

5. Features such as proficiency of chat or speech and participant's interaction and the social networking delineate of the user will improve the performance. Also cyberbullying instances such as racial, sexual etc. need to be considered. Such features

on social networking websites will be more useful in cyberbullying detection. Sentiment mixture models which will identify different meanings of comment and topic-author-community models which identify conversation of users will be helpful in detection of cyberbullying.

6. Features like semantic characteristics of the post content such as vernacular clauses, active or passive voice, and sarcasm or irony are not considered. These peculiarities can help to distinguish various behavioural patterns.

7. Temporal behavior of commenting, incorporating image features by applying image recognition algorithms, mobile sensor data, etc. are not considered which helps to improve the performance of the cyberbullying classifier.

8. Distributed stream and parallel processing can be used to implement real-time detection of cyberbullying. Mediums such as Hadoop, Storm and Flink can be employed to implement the framework to discover cyberbullying.

Various platforms where cyberbullying can occur include online discussion, live conference, gaming platforms and social networking platforms such as Facebook, Twitter, YouTube, Snapchat where people can view and participate in sharing contents such as text, images, audios and videos. Many researchers have worked on the social media datasets because of its open access to the researchers' community.The major dataset sources are Twitter, Formspring, YouTube, Instagram, Vine, Wikipedia, Myspace and Kongregate. Various studies including the work of [1, 7, 8, 9, 10, 11, 13, 14, 17, 19] used dataset of Twitter whereas the work of [3, 10, 13, 23] utilized Formspring dataset. Both datasets contain text in the form posts. YouTube is used by the researchers of [8, 13, 20], where text comments on videos are taken into account. Vine dataset contain videos of 6 seconds long each which is called media session. It is used by [4, 15]. Instagram dataset is used by researches of [4, 6, 12, 16]. Myspace is used by [11, 19, 22]. Authors of [4] have used Instagram and Vine dataset. [2, 10] has used dataset from Wikipedia. Many researchers have worked on multiple datasets such as Twitter, Fromspring and YouTube as in [13] and Formspring, Twitter and Wikipedia as in [10]. Ask.fm and Instagram datasets are used by [12]. Vinita et al. in [19] have worked on Myspace, Slashdot, Kongregate and Twitter datasets. English and Dutch language datasets have been used by Cynthia et al. in [5].

5. Conclusion and Future Directions

Although social media platform has become an important entity for everyone, cyberbullying has many negative impacts on someone's life which includes depression, anxiety, anger, fear, trust issues, low self-esteem, exclusion from social events and sometimes suicidal behaviour too. Cyberbullying instances are not only taking place using texts, but also audio and video features play an important role in spreading cyberbullying. This study has discussed detailed and comprehensive review of the previous research done in the field of cyberbullying. The comparative study has presented different datasets used by researchers. Also methodologies, classifiers and the scope for improvement has also been discussed. This study gives direction to carry out the further research in this field.

Although the significant work has been done in text-based cyberbullying detection, working on audios/ videos cyberbullying detection is still a challenge. Gathering or preparing enough dataset is a major test. Considering social aspects of victim or attacker can be the added input to achieve improved detection of cyberbullying. Detection of cyberbullying can be enhanced if it is able to collaborate with other field such as psychologist and sociologist.

References

[1] Despoina Chatzakou, Ilias Leontiadia, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, Nicolas Kourtellis, "Detecting Cyberbullying and Cyberaggression in Social Media", ACM Transactions on the Web, Vol. 13, No. 3, Article 17, October 2019.

[2] Suresh Mestry, Hargun Singh, Roshan Chauhan, Vishal Bisht, Kaushik Tiwari, "Automation in Social Networking Comments With the Help of Robust fast Text and CNN", 1st International Conference on Innovations in Information and Communication Technology (ICIIACT), June 2019.

[3] John Hani Mounir, Mohamed Nashaat, Mostafaa Ahmed, Eslam A. Amer, "Social Media Cyberbullying Detection using Machine Learning", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, January 2019.

[4] Lu Cheng, Jundong Li, Yasin N. Silva, Deborah L. Hall, Huan Liu, "XBully: Cyberbullying Detection within a Multi-Modal Context", Proceedings of the 12th ACM International Conference on Web Search and Data Mining, WSADM, January 2019, pp. 339-347.

[5] Cynthia Van HeeID, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Ve´roniqueHoste, "Automatic detection of cyberbullying in social media text", PLOS ONE, October 2018.

[6] Kristo Radion Purba1, David Asirvatham, Raja Kumar Murugesan, "A Study on the Methods to Identify and Classify Cyberbullying in Social Media", Proceedings of the Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), October 2018.

[7] Sabina Tomkins, Lise Getoor, Yunfei Chen, Yi Zhang, "A Socio-linguistic Model for Cyberbullying Detection", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASNAM), August 2018.

[8] Hitesh Kumar Sharma, K Kshitiz, Shailendra, "NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms", Proceedings of the International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, June 2018.

[9] Ankita Shekhar, M. Venkatesan, "A Bag-of-Phonetic-Codes Model for CyberBullying Detection in Twitter", Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India, March 2018.

[10] Sweta Agrawal, Amit Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", Springer International Publishing AG, part of Springer Nature, January 2018, pp. 141–153.

[11] Harsh Dani, Jundong Li, and Huan Liu, "Sentiment Informed Cyberbullying Detection in Social Media", Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, December 2017, pp 52-67.

[12] Elaheh Raisi, Bert Huang, "Cyberbullying Detection with Weakly Supervised Machine Learning", Proceedings of the IEEE/ACM International Conference, August 2017.

[13] Michele Di Capua, Emanuel Di Nardo, Alfredo Petrosino, "Unsupervised Cyber Bullying Detection in Social Networks", Proceedings of the 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, México, December 2016.

[14] Vivek K. Singh, Qianjia Huang, Pradeep K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASNAM), August 2016.

[15] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, Sabrina Arredondo Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in Vine", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, August 2015.

[16] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq,Richard Han, QinLv, Shivakant Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network", Association for the Advancement of Artificial Intelligence, March 2015.

[17] Vivek K. Singh, Qianjia Huang, Pradeep K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis", 3rd International Workshop on Socially-Aware Multimedia - SAM, November 2014.

[18] Cicero Nogueira, Dos Santos, MairaGatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts", International Confeence on Computational Linguistics, August 2014.

[19] Vinita Nahar, Xue Li, Chaoyi Pang, Yang Zhang, "Cyberbullying Detection based on Text-Stream Classification", Proceedings of the 11th Australasian Data Mining Conference, Canberra, Australia, November 2013.

[20] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong, "Improving cyberbullying detection with user context", Proceedings of the European Conference on Information Retrieval, March 2013, pp. 693-696.

[21] Vinita Nahar, Sayan Unankard, Xue Li, Caoyi Pang, "Sentiment Analysis for Effective Detection of Cyber Bullying", Proceedings of the 14th Asia-Pacific international conference on Web Technologies and Applications, April 2012.

[22] Maral Dadvar, Franciska de Jong, Roeland Ordelman, Dolf Trieschnigg, "Improved Cyberbullying Detection Using Gender Information", 12th -Dutch-Belgian Information Retrieval Workshop, February 2012, pp. 693–696.

[23] Kelly Reynolds, April Kontostathis, Lynne Edwards, "Using Machine Learning to Detect Cyberbullying", Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops, December 2011.