WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges

***J. Renuga Priyadharshini*[#1], *K.R.Aruna*[#2]**

[#1] M.Sc, Department of Computer Science, Kamban College of Arts and Science for Women, Tiruvannamalai-606603

[#2] Head of the department, Department of Computer Science, Kamban College of Arts and Science for Women, Tiruvannamalai- 606603.

**ABSTRACT:**

The recent developments of communication technologies have considerably transcended the temporal and spatial limitations of traditional communications.These social technologies have created a revolution in user-generated information, online human networks, and rich human behaviour-related data. The fast growing use of social networking sites among the teens have made them vulnerable to get exposed to bullying. Cyberbullying is the use of computers and mobiles for bullying activities. Comments containing abusive words effect psychology of teens and demoralize them. The scourge of cyberbullying has assumed alarming proportions with an ever-increasing number of adolescents admitting to having dealt with it either as a victim or as a bystander. Anonymity and the lack of meaningful supervision in the electronic medium are two factors that have exacerbated this social menace. Comments or posts involving sensitive topics that are personal to an individual are more likely to be internalized by a victim, often resulting in tragic outcomes. Our initial experiments show that using features from our hypotheses in addition to traditional feature extraction techniques like TF – IDF and N – gram increases the accuracy of the system.

## 1.INTRODUCTION:

With the spread of mobile technologies, cyber bullying has become an increasing problem, especially among teenagers. Awareness has also increased, due to some episodes of suicide. According to recent studies almost 43% of teenagers in the U.S. revealed to be victims of cyber bullying. It is, therefore, evident that the availability of tools that can automatically identify possible behaviors classified as cyber bullying, can be really useful to prevent situations of "risk" to the victim. Even if the problem is now heavily considered from a social point of view, computational studies in this field are largely yet unexplored and only few researches on cyber bullying are available. We propose a possible solution for automatic detection of the bully traces, i.e. social media posts containing harmful text or sentence that could possibly lead to a cyber bullyingepisode. We shall show that using both techniques derived from NLP, in the pre-processing data stage, and the subsequent adoption of unsupervised machine learning algorithms, for the detection phase, can lead to reliable results. We propose here a new model of cyber bullying detection

## 2.Literature Survey

[1] 2016**, T Mahlangu**, As we see the cyberspace evolve we also see a directly proportional growth of the people using the cyberspace for communication. As a result, the misuse of the cyberspace has given rise to negative issues such as cyberbullying, which is a form of harassing other people using information technology in a deliberate and continual manner. The detection and prevention of cyberbullying becomes critical for safe and health social media platforms. In this paper, a review of the cyberbullying content in Internet, the categories of cyberbullying, data sources containing cyberbullying data for research, and machine learning techniques for cyberbullying detection are overviewed. The main challenges of the cyberbullying detection are demonstrated, including the lack of multimedia contentbased detection and availability of public accessible dataset. Suggestions are provided as the conclusion of the overview.

[2] 2018, **A Shekar**,Social networking sites such as Twitter, Facebook, MySpace, Instagram are emerging as a strong medium of communication these days. These have become a part and parcel of daily life. People can express their thoughts and activities among their social circle with brings them closer to their community. However this freedom of expression has its drawbacks. Sometimes people show their aggression on Social Media which in turn hurts the sentiments of the targeted victims. Certain forms of cyber-bullying are sexual, racial and physical disability based. Hence a proper surveillance is necessary to tackle such situations. Twitter as a micro-blogging site sees cyber abuse on a daily basis. However, tweets are raw texts; containing a lot of misspelled words and censored words. This paper proposes a novel method to detect cyber-bullying, a Bag-of-Phonetic-Codes model. Using pronunciation of words as features can rectify misspelled words and can identify censored words. Correctly identifying duplicate words can lead to smaller vocabulary of words, thereby reducing the feature space. The inspiration for this proposed work is drawn from the famous Bag-of-

Words model for extracting textual features. Phonetic code generation has been done using the Soundex Algorithm. Besides the proposed model, experiments were carried out with both supervised and unsupervised machine learning approaches on multiple datasets to understand the approaches and challenges in the domain of cyber-bullying detection.

[3] 2018, **H.Rosa,** As cyberbullying becomes more and more frequent in social networks, automatically detecting it and pro-actively acting upon it becomes of the utmost importance. In this work, we study how a recent technique with proven success in similar tasks, Fuzzy Fingerprints, performs when detecting textual cyberbullying in social networks. Despite being commonly treated as binary classification task, we argue that this is in fact a retrieval problem where the only relevant performance is that of retrieving cyberbullying interactions. Experiments show that the Fuzzy Fingerprints slightly outperforms baseline classifiers when tested in a close to real life scenario, where cyberbullying instances are rarer than those without cyberbullying.

[4] 2017, **BatoulHaidar,Maroun, chamounFadiYamout, :** Cyberbullying is the new form of bullying; executed by electronic media and Internet. Cyberbullying is affecting a lot of children around the world including Arab countries. Awareness for cyberbullying is arising and research is taking place in the fields of cyberbullying detection and mitigation and not just the psychological effects of cyberbullying on the victim. Researches on cyberbullying detection have been done in many languages but none has been done on Arabic language cyberbullying detection until the time of writing this paper. Many techniques are utilized in the area of cyberbullying detection' mainly Machine Learning (ML) and Natural Language Processing (NLP). This paper presents a brief background on cyberbullying and all technologies incorporated under this field; in addition to an extensive survey regarding the techniques and advancements in multilingual cyberbullying detection; and finally proposes a plan of a solution for the problem of Arabic cyberbullying.

*[5] BatoulHaidar, MarounChamoun, Ahmed Serhrouchni,* In the era of Internet and electronic devices bullying shifted its place from schools and backyards into the cyberspace; it is now known as Cyberbullying. Children of the Arab countries are suffering from cyberbullying same as children worldwide. Thus concerns from cyberbullying are elevating. A lot of research is done for the purpose of handling this situation. The current research is focusing on detection and mitigation of cyberbullying; while previous research dealt with the psychological effects of cyberbullying on the victim and the predator. A lot of research proposed solutions for detecting cyberbullying in English language and a few more languages, but none till now covered cyberbullying in Arabic language. Several techniques contribute in cyberbullying detection, mainly Machine Learning (ML) and Natural Language Processing (NLP). This paper presents a solution for detecting and stopping cyberbullying with focus on content written
in Arabic Language. Thus the primary results of the system are displayed and discussed.

## 2.1PROPOSED SYSTEM

Since it is challenging though interesting to capture the unobserved links between existing users, instead of analyzing the overall social network, our group decides to focus on individuals and predict the plausible link creations of a given user. To achieve this goal, we want to find algorithms that will utilize information of the network structure as well as node attributes.

With the models built based on our chosen algorithms, we can predict future link creations among SOCIALMEDIA users and provide them with a list of people that they are likely to follow.

## 3. METHODOLOGY

### 3.1 **Background**

Many real world interactions are richly structured. Entities of different types are related to each other through networks.

Social network, like SOCIALMEDIA, is a complex, dynamic and noisy network system due to its nature of constantly changing interactions (i.e. new connections among members). Predicting the occurrence of links between users is an important issue to draw attentions on in such networks.

Link prediction can be used to suggest most likelymatch users' interests and needs, a social network will be able to improve its user loyalty and maximize the user experience.

### 3.2 **Data Gathering**

the connections among users. The data set we are going to use in our project is the 1% sample stream from SOCIALMEDIA. We are going to extract related attributes of users and list of people that each user follows from the given dataset. The filtered data will be used in the two algorithms that will be addressed in the next section. We will introduce directed graph to illustrated the connections users. link creation in the future. It can also be used to predict unobserved links in a network to indicate the close relationship among individuals even though their interactions

Need data from In the emo-net corpus, we extracted four sets of tweets in the English language according to the following search criteria: a) tweets associated to immigrant and war related events (e.g. terrorist, terrorism, ISIS, etc.); b) tweets containing negatively polarized words (e.g. anger, fear, c) tweets associated to pets (e.g. puppy, kitty, etc.) and d) tweets containing positively polarized words (e.g. joy, happiness, happy, etc.). We will refer to the networks constructed from these sets of tweets respectively as: a) emo-neta , b) emo-netb , c) emo-netc and d) emo-netd . The four search criteria are selected in order to ensure consistency with the positively or negatively annotated polarity of tweets in the SC dataset, and to keep the data used for the experimental set-up comparable.

### 3.3 Model and Methodology

First of all, define a user from SOCIALMEDIA as a node, named it u. To restate our problem, it is to find the best link predictions and provide link recommendations for u. We decide to approach our problem using two algorithms and combine them in a way that will hopefully outperform both algorithms. The models selected and methodologies used will be discussed separately for these two algorithms. The first algorithm we are going to use is a classification approach. In another word, we want to design an algorithm that inputs selected attributes for a given u and learns to distinguish “positive” nodes (links that will be created in the future) from “negative” nodes (links that will not be created) in the given training data.

### 3.4 Sentence Offensiveness

Prediction Offensive messages always include offensive words. Strongly offensive words, such as “f***” and “s***”, are conventionally and generally offensive; but other weaker offensive words, such as “stupid” and “liar”, are less identified. The study differentiates between these two types of offensive words, and assigns offensiveness levels accordingly. Hence, the definitions of offensiveness level of each pejorative, w, in sentence, s Once a pejorative describes an online user, or semantically associates with another pejorative, it becomes more offensive from users’ perceptions. Thus, an intensifier is needed to scale the offensiveness value of words. We use typed dependency parser, proposed by Stanford Natural Language Processing Group2 , to capture the grammatical dependencies within a sentence. If sentence, s, is parsed and all words semantically related to pejorative, w, are categorized as set { , } , the intensifier, Iw, of word, w, can be defined as:

$$I_w = \begin{cases} b_1 & \text{if } \exists d_i \in D_{w,s},\ d_i \text{ is a user identifier} \\ b_2 & \text{if } \exists d_i \in D_{w,s},\ d_i \text{ is an offensive word} \\ 1 & \text{otherwise} \end{cases}$$

### 3.5 User Offensiveness Aggregation

Synthesizing the offensiveness values of all sentences allows computing the overall offensiveness of the user. Thus, given a post, p, containing sentences, {s1,…sn} , and the offensiveness of the sentences, { , } 1 n s s o Λo , the offensiveness, Op , of p should be, Op =∑os . Hence, the offensiveness value, Ou, of a user who has m posts is, Ou = m∑Op 1 , because users who frequently post offensive messages are more offensive than occasional offenders. If Ou>0, then the conversation history of user, u, indicates offensiveness to some extent

Bag-of-words (BoW) approach: The BoW approach disregards grammar and word order and detects offensive sentences by checking whether or not they contain user identifiers and offensive words.

N-gram approach: The N-gram approach detects offensive sentences by selecting all sequences of n-words in a given sentence and checks whether or not the sequences include user identifiers and offensive words.

Appraisal approach: The Appraisal approach detects offensive sentences by checking whether or not phrases in a given sentence direct certain offensive words towards an online user.

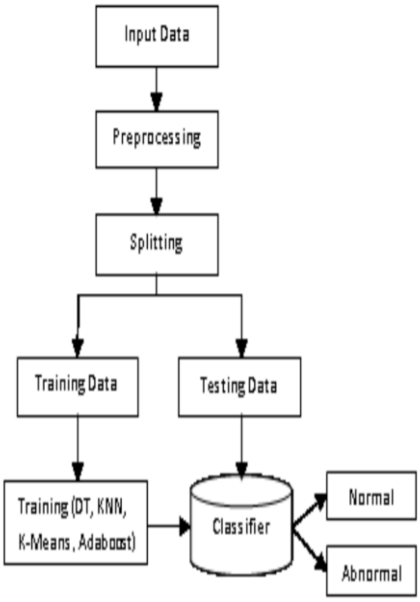
### 3.6 Evaluation of various Feature Sets

Most of the previous studies to detect phishing have used features based on the URL of the suspicious page and the HTML source of the landing page. In this study, we propose to use SOCIALMEDIA based features along with URL based features to quickly detect phishing on SOCIALMEDIA at zero-hour. To evaluate the performance of detection using these additional set of features based on SOCIALMEDIA properties, we present feature-set wise performance of the classification technique we use. As described in Table I, we have used four sets of features in this study. To evaluate the impact of each feature set, we performed classification task by taking one feature set at a time and then added the other one in the next iteration. Table V presents our experiment results by using different set of features using Random Forest classification method which gives us the overall highest accuracy of 92.52%. We observe that when we use only URL based features, we get an overall accuracy of 82.22% and a low precision and recall for ‘phishing’ class. The addition of SOCIALMEDIA based feature sets, user based features and network based features significantly improve the performance of phishing detection and boost the precision of identifying phishing tweets significantly. Hence, SOCIALMEDIA based features are helpful in increasing the performance of classifying phishing tweets.

### 3.7Parallel computation of features:

To enable quick decision on a tweet, we have multiprocessing modules in our system which extract features in parallel. This helps in reduction of overall computation time. However, in future, we can further improve the feature extraction by distributing the computation of features across multiple servers.

## 4. SYSTEM ARCHITECTURE:



## 5. CONCLUSION:

This examination looked into existing writing to identify forceful conduct on SM sites by utilizing AI draws near. We explicitly inspected four parts of distinguishing cyber bullying messages by utilizing AI draws near, to be specific, information assortment, include building, development of cyber bullying identification model, and assessment of built cyber bullying identification models. A few sorts of discriminative highlights that were utilized to identify cyber bullying in online long range informal communication locales were likewise abridged.

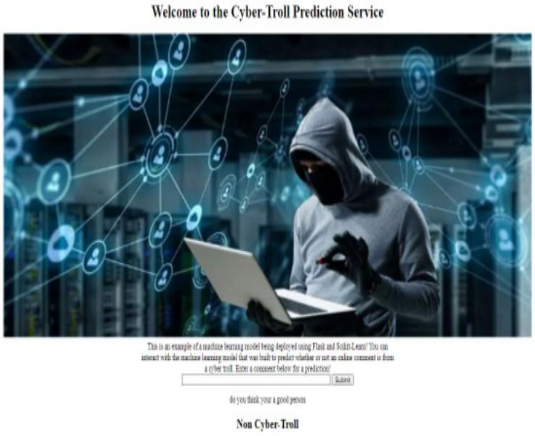
Moreover, the best regulated AI classifiers for characterizing cyber bullying messages in online long range informal communication locales were distinguished.

One of the fundamental commitments of current paper is the meaning of assessment measurements to effectively distinguish the critical parameter so the different

AI calculations can be assessed against one another. In particular we abridged and distinguished the significant components for identifying cyber bullying through AI methods exceptionally directed learning. For this reason, we have utilized precision, accuracy review and f-measure which gives us the region under the bend work for displaying the practices in cyber bullying.

## RESULTS:

**Fig 1: Detection of Cyber Bullying Words**



Welcome to the Cyber-Troll Prediction Service

This is an example of a machine learning model being deployed using Flask and Scikit-Learn! You can interact with the machine learning model that was built to predict whether or not an online comment is from a cyber troll. Enter a comment below for a prediction!

do you think your a good person

**Non Cyber-Troll**

## REFERENCES

[1] V. Subrahmanian and S. Kumar, ''Predicting human behavior: The next frontiers,'' Science, vol. 355, no. 6324, p. 489, 2017.

[2] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, ''Homophily in the digital world: A LiveJournal case study,'' IEEE Internet Comput., vol. 14, no. 2, pp. 15–23, Mar./Apr. 2010.

[3] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, ''Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network,'' Comput. Hum. Behav., vol. 63, pp. 433–443, Oct. 2016.

[4] L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, ''Using social media to predict the future: A systematic literature review,'' 2017, arXiv:1706.06134. [Online]. Available: https://arxiv.org/abs/1706.06134

[5] H. Quan, J. Wu, and Y. Shi, ''Online social networks & social network services: A technical survey,'' in Pervasive Communication Handbook. Boca Raton, FL, USA: CRC Press, 2011, p. 4.

[6] J. K. Peterson and J. Densley, ''Is social media a gang? Toward a selection, facilitation, or enhancement explanation of cyber violence,'' Aggression Violent Behav., 2016.

[7] BBC. (2012). Huge Rise in Social Media. [Online]. Available: http://www.bbc.com/news/uk-20851797

[8] P. A. Watters and N. Phair, ''Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA),'' in Cyberspace Safety and Security. Berlin, Germany: Springer, 2012, pp. 66–76.

[9] M. Fire, R. Goldschmidt, and Y. Elovici, ''Online social networks: Threats and solutions,'' IEEE Commun. Surveys Tuts., vol. 16, no. 4, pp. 2019–2036, 4th Quart., 2014.

[10] N. M. Shekokar and K. B. Kansara, ''Security against sybil attack in social network,'' in Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES), 2016, pp. 1–5.

[11] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, ''Detecting and tracking political abuse in social media,'' in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 297–304.

[12] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ''PhishAri: Automatic realtime phishing detection on Twitter,'' in Proc. eCrime Res. Summit (eCrime), Oct. 2012, pp. 1–12.

[13] S. Yardi et al., ''Detecting spam in a Twitter network,'' First Monday, Jan. 2009. [Online]. Available: https://firstmonday.org/ article/view/2793/2431

[14] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, ''Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter,'' in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 71–80.

[15] G. R. S. Weir, F. Toolan, and D. Smeed, ''The threats of social networking: Old wine in new bottles?'' Inf. Secur. Tech. Rep., vol. 16, no. 2, pp. 38–43, 2011.

[16] M. J. Magro, ''A review of social media use in e-government,'' Administ. Sci., vol. 2, no. 2, pp. 148–161, 2012.

[17] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ''Improving cyberbullying detection with user context,'' in Advances in Information Retrieval. Berlin, Germany: Springer, 2013, pp. 693–696.

[18] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, ''Detecting offensive language in social media to protect adolescent online safety,'' in Proc. Int. Conf. Privacy, Secur., Risk Trust (PASSAT), Sep. 2012, pp. 71–80.

[19] V. S. Chavan and S. S. Shylaja, ''Machine learning approach for detection of cyber-aggressive comments by peers on social media network,'' in Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI), Aug. 2015, pp. 2354–2358.

[20] W. Dong, S. S. Liao, Y. Xu, and X. Feng, ''Leading effect of social media for financial fraud disclosure: A text mining based analytics,'' in Proc. AMCIS, San Diego, CA, USA, 2016.