Content-based Cybercrime Detection: A Concise Review

Amanpreet Singh, Maninder Kaur

Abstract: ***In the recent past, the issues of Content-based Cybercrime have gained considerable attention. Social media providers seek for accurate and efficient way of recognizing offensive content for shielding their users. Content-based Cybercrime detection is one of the conspicuous area of data mining that deals with the recognition and examination of bully contents usually presented at social media. The current work emphasizes on cyberbullying, one of the prominent problems that arose due to the increasing fame of social network and its fast acceptance in our day-to-day survives. The social network provides a convenient platform for the cyber predators to bull their preys especially targeting young youth. In severe cases, the victims have attempted suicide due to humiliation, insult, and hostile messages left by the predators. This work presents a systematic critical study to accumulate, investigate, apprehend and explore the patterns and study gaps in a well-organized manner. The study portrays a comprehensive systematic literature review of strategies proposed in the field of content-based cybercrime. In this review, precise investigation methodology is utilized based on a total selected 27 research papers out of 51 research papers published in preeminent workshops, symposiums and conferences and conspicuous journals. The survey relates to several data pre-processing techniques, content-based feature, machine learning methodology, online social networking datasets and evaluation parameter used in context of detecting content-based cybercrime. This Methodical analysis of the research work acts as an assistant for the researchers to discover the significant characteristics of content-based Cybercrime detection techniques.***

Index Terms: Content based cybercrime, cyberbullying, machine learning, deep learning.

I. INTRODUCTION

Internet has changed majority aspects of human lifetime: education, entertainment, politics, relationships and so on. It affects someone’s mood: they feel connected, happy, loved, lonely, depressed, scared and so forth. Maybe not willingly, but undoubtedly our lives have become interwoven with Internet. Social networking provides the humanity a huge and convenient platform for exchanging their ideas, perceptions all around the world. There is a continuous growth of the size of social networks. Figure I depict the count of active users (in millions) involved in social networking websites [1]. A thousand millions of users throughout the globe are using one or more social networking sites, with the count increasing rapidly at fast pace. This count includes every age group whether young or adult and both males and females. Out of this huge count, nearly millions of the individuals are habituated to these social networking sites. At the present time, the friendships and relationship networks are shaped through a wide array of digital devices. The majority of daily greetings, friendly get-togethers and family chitchats take place from behind a screen. Most of the time people reach out to others for help, love and friendship, but on the other side, hostility and hatred have also always been part of human culture and they have detrimental impact on societal history. Besides convenience and extreme openness, social networking platform can be effortlessly utilized for spreading uncivilized and unsocial activities. The offensive wrong doings and patterns of behavior driven by the darker sides of human nature can be observed in these virtual settings. The social media prompt the youth into a globe of disastrous threats such as Cybercrime.

Cybercrime is viewed as one of the most hazardous fears for the expansion of any vulnerable situation; it has a severe influence on every facet of the development of a state. Administration bodies, non-profit governments, remote industries and residents are all probable targets of the cyber-criminal crowd. The "cybercrime industry" operates specifically as authentic administrations working on an international level, with safety experts approximating the common measure of misfortunes to be processed in the demand of billions of bucks each year. The term ‘Cybercrime’ is stated as any illegitimate action that uses computer as the prime mechanism of committing crime. The U.S. Department of Justice extended this definition: “For any illegal activity, use a computer as storage of evidence” [2].

Cyber-crime is classified in two categories: technology based and content-based crime. Any particular terrorist group associated to sexual harassment, fear, child pornography, national security etc. accomplishes the content-based crime. The technology-based cybercrime includes hacking, incidents of espionage, injecting malicious code [3]. Figure II shows the taxonomy of cybercrime with some examples. The folks tangled in both categories should have some skilled consciousness. The cyber crooks generally inclined to live in different categories of globe and relish receiving the honor of several nations.

Revised Manuscript Received on June 05, 2019

Amanpreet Singh, Department of Computer Science & Engineering, Thapar Institute of Engineering and Technology, Patiala, India,

Maninder Kaur, Department of Computer Science & Engineering, Thapar Institute of Engineering and Technology, Patiala, India.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

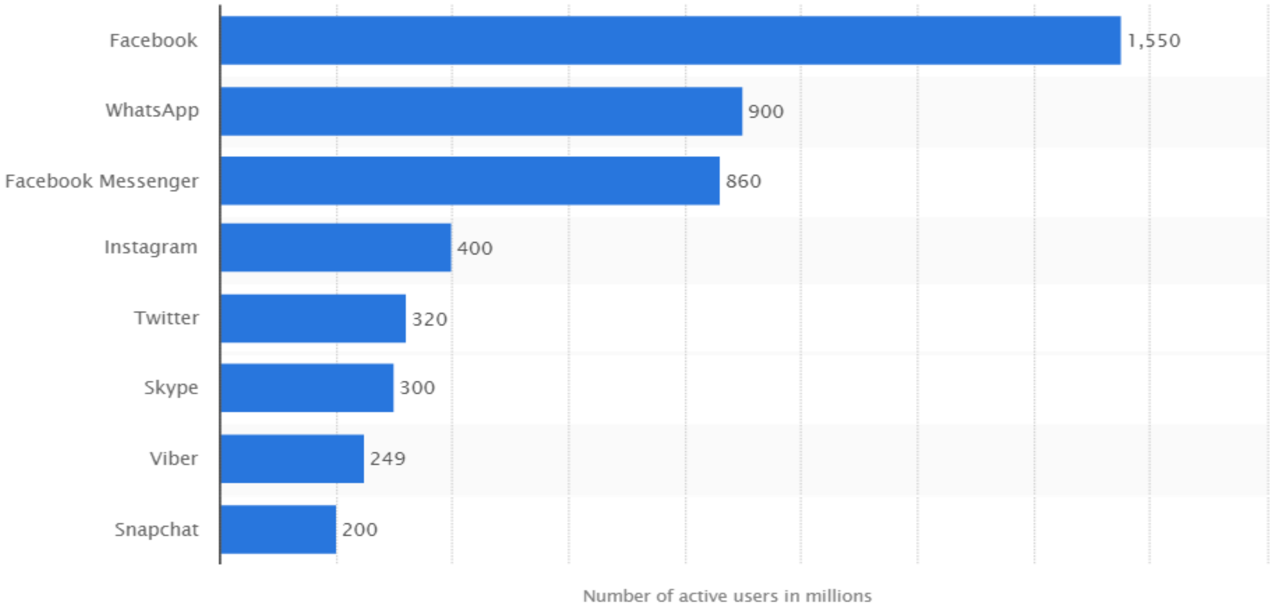

Fig. I: Number of active users (in millions) on various social networking websites [1]

Out of these cybercrimes, content-based cybercrime has been a significant and dire issue ever since the emergence of the Internet. One of the main content-based cybercrime problems is cyberbullying where the target subjects are under-age victims. Specifically, cyberbullying has risen as a noteworthy issue alongside the ongoing improvement of online correspondence and social network.

There has been an evident number of life-threatening experiences due to cyberbullying especially among youths throughout the world [4]. The recent studies show that this problem is more exaggerated in the USA where about 43% of teens are the targets of cyber bullying [5]. It is consequently apparent that the readiness of tools that can distinguish potential practices categorized as cyberbullying can be extremely valuable to avert circumstances of “threat” to the prey. Regardless of whether the issue is presently vigorously reflected from a social perspective, computational examinations in this field are to a great extent yet unexplored and just couple of explores on cyber bullying are available.

There is a critical necessity to dig into cyberbullying in context of its prevention, mitigation and detection. The current study focuses on detailed literature review of detection of content-based crime comprising cyberbullying.

A. Background and Motivation

Cyberbullying is sufficiently threatening and destructive that it can lead victims to contemplate suicide and, in the worst scenarios, it can actually result in suicidal attempts [6] and cause life-long mental damage to victims. Numerous deadly cyberbullying encounters have been accounted for globally, consequently drawing consideration towards its negative effect.

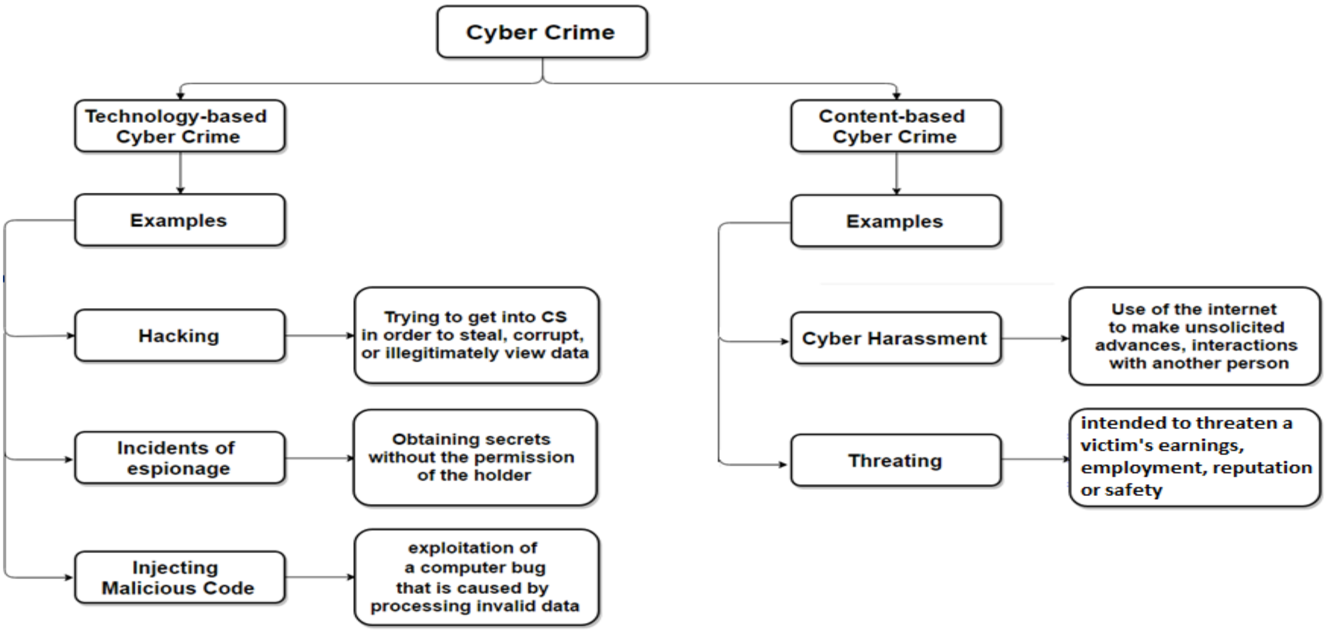
Cyber Crime
Technology-based Cyber Crime
Content-based Cyber Crime
Examples
Examples
Hacking
Trying to get into CS in order to steal, corrupt, or illegitimately view data
Incidents of espionage
Obtaining secrets without the permission of the holder
Injecting Malicious Code
exploitation of a computer bug that is caused by processing invalid data
Cyber Harassment
Use of the internet to make unsolicited advances, interactions with another person
Threating
intended to threaten a victim's earnings, employment, reputation or safety

Fig. II: Taxonomy of cybercrime with examples

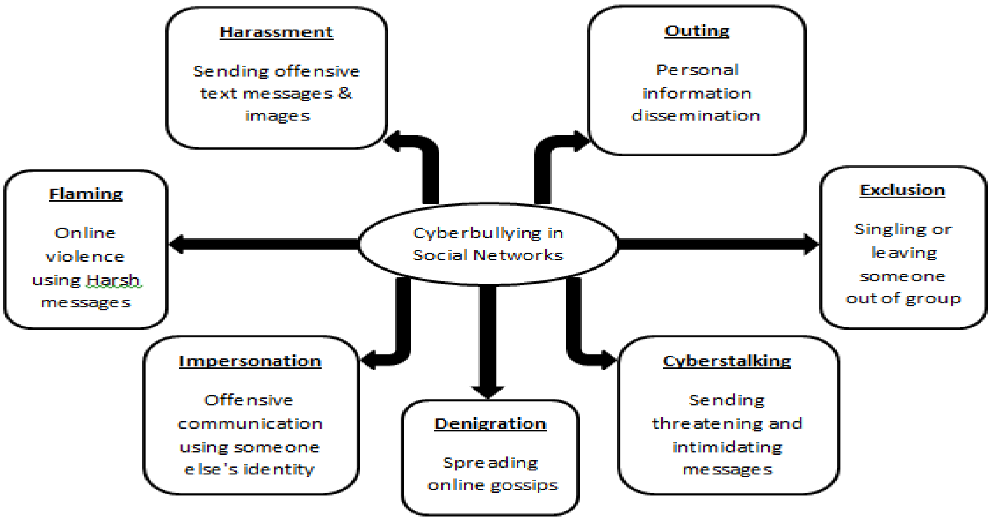
Harassment
Sending offensive text messages & images
Outing
Personal information dissemination
Flaming
Online violence using Harsh messages
Cyberbullying in Social Networks
Exclusion
Singling or leaving someone out of group
Impersonation
Offensive communication using someone else's identity
Denigration
Spreading online gossips
Cyberstalking
Sending threatening and intimidating messages

Fig. III. Various ways of cyberbullying in online social networks

Academics have recognised the importance of the early detection of such activities. However, existing work in this area has been primarily focused on determining the presence of cyberbullying based on textual information and supervised learning.

We identified the need of systematic literature review after taking into account revolutionary research in content-based cybercrime detection. The work summarizes the available research on the basis of broad and systematic search in prevailing repository.

II. CYBERBULLYING

Cyberbullying is one form of online misbehavior that has deeply affected society with harmful consequences. Traditional bullying used to be a demonstration of dominance and consolidation of social status by making use of physical power and creating fear and discomfort for those who were weaker and vulnerable. With the development of online technology, bullying has also emerged in cyber societies. Cyberbullying can simply be stated as an intentional action that is conducted through digital technology to hurt someone [7]. Unlike traditional bullying, which was inherently limited to streets and school yards, the vast variety of technological devices used in daily lives has brought cyberbullying also into people's homes and bed rooms.

Bullying is usually defined as a subcategory of aggressive behavior [8]. It is characterized by reiteration over period and a disparity of authority between bully and target [7, 8]. There are different categories of common cyberbullying [9,10] as shown in Figure III:

- ***Flaming***: Sending rude and vulgar messages to a group or person.
- ***Outing***: Posting private information (picture, phone number…) or manipulated/photo-shopped personal materials of an individual without her or his consent.
- ***Harassment***: Repeatedly sending insulting messages or emails to a person.
- ***Exclusion***: Excluding someone from participating in an online group.
- ***Impersonation***: Fantasizing to be another person for directing out materials on her or his behalf.
- ***Cyberstalking***: Terrorizing someone by sending threatening and intimidating messages.
- ***Denigration***: Spreading online gossips about a person.

Cyberbullying can happen through different modalities. It can happen through posting nasty videos about someone or publicly uploading private pictures without having the consent of their owner. Cyberbullying through text is one of the most common mediums, in which vulgar comments are posted and threatening and foul messages are sent to the victim.

There are various algorithms and tools that can automatically detect and remove bullying posts, or trigger some kind of an administrator's follow-up action in response to online bullying incidents.

A. Components of Cyberbullying

Cyberbullying consists of several components. These components affect how the bullying takes place and consequently the studies conducted on cyberbullying differ depending on the components involved.

- The fundamental component is the people, called ***actors***, involved in the incident. The actors can be grouped into the following three categories:
  - Bully: the person who intentionally uses obscenity, threat or aggression to impose domination or cause fear and distress in others.
  - Victim: the person who is targeted by the bully. Victims cannot easily defend themselves and are usually vulnerable to the imbalance of power between them and the bully.
  - Bystander: the person who witnesses the incident but is not
  - directly involved in the process. The bystanders can provide support for the
  - victim by posting positive feedbacks for the victim

and reacting against the bullies. They can also escalate the distress caused by the bullies,

- by supporting their actions.

Table I: Cyberbullying components in pre- and post-bullying phases and the actions that could be triggered by the prediction modules proposed.

		Pre-Bullying		Post-Bullying
Actors	Bully	To be monitored	Bullying	To be identified/ To be warned or to be excluded from the network
	Victim	To be trained To be educated		To be identified/ To receive support
Bystanders		To be alerted To be monitored		To be alerted/ To be monitored
Platform		Exclusion of risky user profiles		Identification of bullies and victims. Follow-up actions, e.g., organizing help after incident, alerting of bystanders, removing offensive
Content		Previously analyzed content to be used to identify risky user profiles		Bullying content to be detected, offensive content to be deleted

- The platform in which cyberbullying takes place is another influential component in the process and therefore it should also be taken into consideration in the studies.
- Another component is the content and the modality through which the bullying takes place. As explained earlier, cyberbullying can happen through videos, pictures as well as through posting hurtful and offensive textual contents.

In the study of measures addressing the pre-bullying phase, the main concentration is on prevention strategies while in the study of measures addressing the post-bullying phase the focus is on the detection of bullying incidents after they have happened. Table I illustrates the status of the components distinguished in each phase.

III. EXISTING STUDIES

This study includes research efforts on content-based cybercrime detection in online social networks. The review incorporates articles, printed over the last decade, beginning through the innovative effort of Yin et al. 2009. The extent of the work incorporated in the review highlights the increasing concern received in recent years by cyberbullying prevention. Although the methods considered by many studies are dominated by supervised learning approaches, scholars have shown readiness to use emergent effort from different fields of Natural Language Processing (NLP) to improve the performance.

We discuss former studies from eight perspectives that are: used methodology, conclusions or findings, demerits, dataset used, preprocessing steps used, content-based features used, models or technique, and evaluation metrics used as illustrated in table II and table III.

For the purpose of searching, electronic literature is explored through Scopus, IEEE Xplore virtual library and the ACM digital Library. The key search method was related to the subject “detection of content-based cybercrime, unsocial behavior and harassment” without considering publication year as a filter.

More papers were discovered using the citations in observed articles via the article’s references as an initial topic, thereby discovering 51 academic papers in total as a result of the search. At the very first stage, the titles, abstract, and concluding arguments of the discovered papers were reviewed for assessing their relevance and 18 papers were discarded, as they were not found to be applicable to the survey. In the next phase, the detailed review of whole text of the remaining 33 articles was carried out and further six more papers were filtered out that did not focus on cyberbullying detection. This process led to 27 papers in the final list of papers for the purpose of survey. These papers under consideration dealt with different subjects such as story coordinating to recognize upset teenagers, youth violence participation recognition, cyberbully inhibition strategies [11-37].

The survey targeted on the abstract view of the work done in terms of data sources, availability of the datasets, detection techniques, features extracted, evaluation parameter used, and pre-processing steps.

Tables II presents a summary of key statistics abstracted from the reviewed research. It delivers a concise outline of methodology used, conclusion/findings, demerits and dataset used, types and recognition tasks for each of the 27 papers. It is revealed from the survey that the most common task executed in cyberbullying detection is the binary classification. In this context, the text containing bully terms are entitled as the member of “bullying” class and the message without bully terms categorized as “non bully” type. The significant job after this is the recognition of documents that own the essential features of the “bullying” type. The 23 research papers out of 27 research reviewed, targeted binary classification solely as a cyberbullying detection task or as a hybrid with other tasks. Herein, the text classification is generally expedited by supervised learning systems.

TABLE. II: Various methodologies used for detecting Content-based Cybercrime

Author(s)	Year	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Yin et al. [11]	2009	Used supervised machine learning approach in which features (local, contextual, and textual) of Documents are utilized to learn an SVM Classifier.	Addition of the sentiment and contextual features provide significant performance to basic model which uses only Local feature.	Dataset is of stand-alone posts, pragmatics of conversation are not considered, only for supervised learning techniques. Predators and victims were not identified.	Datasets of Kongregate, Slashdot, MySpace websites *http://caw3.barcelonamedia.org/*
Dinakar et al. [12]	2011	Used supervised machine learning approach in which binary & multiclass Classifiers classify bullying sensitive topics.	Label-specific (binary) classifiers are more effective than multiclass classifiers at detecting content-based cybercrime.	Dataset is of stand-alone posts, pragmatics of conversation are not considered, only for supervised learning techniques. Predators and victims were not identified.	Youtube comments from different videos after clustering into Physical appearance, sexuality, race & culture.
Bayzick et al. [13]	2011	Proposed a rule-based system called BullyTracer and also developed a truth-set of MySpace threads to check accuracy of proposed system.	Correctly identify 85.3% as cyberbullying posts and 51.91% as innocent posts of MySpace dataset.	Falsely flag a lot of innocent posts as cyberbullying. Only uses rule-based system, no supervised or unsupervised learning technique was used.	Thread-style forum transcripts crawled from MySpace media. Link: MySpace.com
Reynolds et al. [14]	2011	Supervised machine learning approach in conjunction with labelled data was used to learn the system to identify bullying content.	Model was capable to recognize 78.5% posts in Formspring dataset that have cyberbullying in a small sized sample.	Only for Supervised Learning techniques, pragmatics of conversation are not considered, dataset is of stand-alone posts. Predators and victims were not identified.	18554 user's data which contain 1 to 1000 posts is used. Link: *www.Formspring.me*
Dinakar et al. [15]	2012	Used common-sense knowledge base with associated reasoning techniques in addition to machine learning classifiers.	In the task of detection of textual cyberbullying, binary classifiers outperform multiclass classifiers.	Other aspects of the problem like pragmatics of conversation and dialogue did not considered by the authors.	Manually Labelled corpus of Youtube and Formspring data. Link: *www.Formspring.me*
Nahar et al. [16]	2012	Proposed a sentimental analysis technique for cyberbullying content detection by using PLSA (a method of feature selection).	Finds the Most Influential persons (predator or Victims) using HITS Algorithm.	Not focused on Indirect Cyberbullying.	Datasets of Kongregate, Slashdot, MySpace websites Link: www.*caw3.barcelonamedia.org*
Nahar et al. [17]	2013	Proposed a session-based framework which incorporated an ensemble of one-class classifier and addressed the real-world scenario where just minimal set of positive instances were given.	Effectively classifies bully instances using session-based one-class ensemble classifier which uses small set of labelled data and huge unlabelled data.	Baseline swear-keywords method can be incorporated along to improve the accuracy.	Datasets of Twitter, MySpace, Kongregate, and Slashdot websites Link: *http://caw2.barcelonamedia.org*

Dadvar et al. [18]	2013	Proposed effective technique which used a combination of user-based, content-based and cyberbullying-specific features.	Evaluation Parameter gives best result when all features used in combination.	Some features such as gender, user profile or channel subscribed can be taken into account.	Youtube comments on three top videos for variety of topics.
Nahar et al. [19]	2014	Proposed semi-supervised learning approach using fuzzy SVM classifier.	This technique was suitable in real-world situation handling noisy, imbalanced or streaming data and outperformed all other methods.	Severity levels of cyberbullying messages were not taken into consideration. Also feature space used is static.	Datasets of Kongregate, Slashdot, MySpace websites Link: *http://caw3.barcelonamedia.org*
Nahar et al. [20]	2014	Proposed semi-supervised learning in the session-based framework that incorporates an ensemble of one-class classifiers.	Results indicated that in real world situations, the proposed approach performed very well, where for initial training only a few positive instances of cyberbullying are available.	Not focused on Indirect Cyberbullying, user-based features are not used.	Datasets of Kongregate, Slashdot, MySpace and Twitter websites Link: *http://caw3.barcelonamedia.org*
Huang et al. [21]	2014	Proposed a technique by integrating social network features with the textual features to detect cyberbullying.	Outcomes showed that new attributes (social) are beneficial in identifying cyberbullying. Proposed model also detects the most influential persons.	Not focused on Indirect Cyberbullying, Only for Supervised Learning techniques.	Datasets of Twitter website Link: http://twitter.com
Mangaonkar et al. [22]	2015	Proposed collaborative paradigm that used different machine learning techniques for classification of bully or non-bully data.	Without much tuning of algorithms, collaboration techniques worked better than sequential methodology in terms of time consumed and accuracy.	Lack Very less features are used to train machine learning algorithm.	Datasets of Twitter website Link: http://twitter.com
Hee. et al. [23]	2015	Presented the annotation and construction of a corpus of posts from Dutch social media (ASKfm) and explored the feasibility of automatic cyberbullying detection.	By exploring the automatic cyberbullying detection model's feasibility, the results presented that when more fine-grained categories are taken into consideration the detection of cyberbullying is not a trivial task.	Some features such as syntactic patterns, semantic information can be taken into account. Also, author role information can improve in cyberbullying detection.	Datasets of ASKfm website Link: http://ask.fm
Hosseinmardi et al. [24]	2015	Collected an Instagram data set sample consisting of comments associated with images, and developed a tagging study for image content in addition to cyberbullying using human labellers at the crowd-sourced Websites (like Crowdflower).	Proposed model identified that there is significant class of media sessions of Instagram that exhibits cyber aggression but not cyberbullying.	More features and detailed labelling surveys can improve accuracy.	Datasets of Instagram website Link: http://Instagram.com
Al-garadi et al. [25]	2016	Suggested a feature-based classifier for detecting cyberbullying using supervised machine learning in the Twitter media.	SMOTE + Random forest using proposed features, showed the best results in detecting cyberbullying.	Other social media data and social networking graph can be used to investigate cyberbullying behaviour.	Datasets of Twitter website Link: http://twitter.com

Galán-García et al. [26]	2016	Proposed a methodology for detecting forged Twitter profiles and offered an efficacious real-world use case.	PolyKernel SMO model outperformed in terms of AUC.	More NLP techniques can be used to improve the accuracy, further data from other social media can be used.	Datasets of Twitter website Link: http://twitter.com
Singh et al. [27]	2016	Proposed a framework (probabilistic information fusion) that utilizes interdependencies associated with different textual and social features, their confidence score, and uses those for better cyberbullying predictors.	Proposed fusion approach provides better results in detecting cyberbullying using heterogeneous textual and social features.	Only emphases on a specific social network (Twitter), more sophisticated features can be used.	Datasets of Twitter website Link: http://twitter.com
Zhao et al. [28]	2016	A learning method was proposed for detection of cyberbullying by concatenating bullying, latent semantic and BoW features together.	Capture Semantic Information behind words. Linear SVM is capable of detecting text containing bullying.	Dataset is of stand-alone posts, Only for Supervised Learning techniques.	Datasets of Twitter website Link: *http://research.cs.wisc.edu/bullying/data.html*
Dani et al. [29]	2017	Proposed a framework based on sparse learning by integrating user-post relationships and sentiment information.	Experimental results showed the impact of sentiment information on two real-world datasets as well as effectiveness of the proposed model.	Other languages can be incorporated as well. The effect of the sarcasm facts concealed in the posts can be investigated.	Datasets of Twitter and Myspace website Link: http://twitter.com Link: MySpace.com
Raisi et al. [30]	2017	Proposed a weakly supervised participant vocabulary consistency (PVC) model using machine learning model for simultaneously inferring new vocabulary indicators of bullying and user roles in molestation-based bullying.	Proposed model was analysed on datasets from diverse social media based on qualitative and quantitative evaluation.	Network features or the sequence of conversations can be considered to improve accuracy.	Datasets of Ask.fm, Instagram, and Twitter website Link:
Singh et al. [31]	2017	Designed a predictive classifier based visual features (e.g. portrayed emotions, nudity, race, gender etc.) for automatic cyberbullying detection.	The usage of visual features outperformed textual features in detecting cyberbullying by improving accuracy.	Huge extent of pictorial processing APIs in addition to visual content can be considered for better detection of cyberbullying.	Datasets of Instagram website Link: http://Instagram.com
Zhao et al. [32]	2017	The authors used semantic extension of deep learning model stacked denoising autoencoder for developing Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA).	Proposed method was capable of utilizing the concealed attributes from the bully data and train the distinguishing and strong illustration of text.	By considering order of words in messages could improve the robustness of learning model.	Datasets of Twitter, MySpace websites
Agrawal et al. [33]	2018	Developed four models based on DNN (i.e. BLSTM, LSTM, CNN and BLSTM) for cyberbullying detection in online social sites.	Proposed model systematically analyses detection of cyberbullying based on different themes through numerous SMPs by means of transfer learning and deep learning-based classifier.	Additional information such as data about the users' social graph and their profile could further improve model performance.	Datasets of Formspring, Twitter and Wikipedia websites

Dadvar et al. [34]	2018	Proposed deep learning-based models also evaluated and transferred the model's performance trained on one platform to another platform.	The proposed models based on deep learning outperform the models based on machine learning models by applying on Wikipedia, Twitter, Formspring and YouTube dataset.	Profile info of the social sites' users can also improve the model's accuracy.	Datasets of Wikipedia, Twitter, Formspring and YouTube
Rafiq et al. [35]	2018	Developed a cyberbullying detection system for media-based social networks, consisting of a dynamic priority scheduler, a novel incremental classifier, and an initial predictor.	The proposed system drastically reduces the time to raise alerts and the classification time. Without sacrificing accuracy, it was very receptive in raising alarms and greatly scalable.	Investigate the plateauing effect that limits the effectiveness of adding more memory.	Datasets of Vine
Van Hee et al. [36]	2018	Proposed an automatic detection model for cyberbullying in social sites by modelling texts written by bystanders' preys, and bullies of online bullying.	Experiments reveal that proposed method was a promising approach for detecting signals of cyberbullying automatically in social sites data and outperforms a word n-gram and keyword-based baseline.	Fine-grained categories related to cyberbullying such as hate, racism expressions, curses and threats can be detected.	Datasets of ASKfm Posts in English and Dutch
Cheng et al. [37]	2019	Proposed XBully, a framework for detecting cyberbullying, that initially re articulates multi-modal data from social network and then targets to train node-embedding illustrations upon it.	Broad experimental outcomes on real-life datasets validate the efficacy of the proposed framework and find that multi-modal data can suggest valuable visions for depicting and detecting cyberbullying behaviours.	Building a deeper understanding of various modalities in depicting cyberbullying behaviours will advance detection of cyberbullying.	Datasets of Instagram and Vine

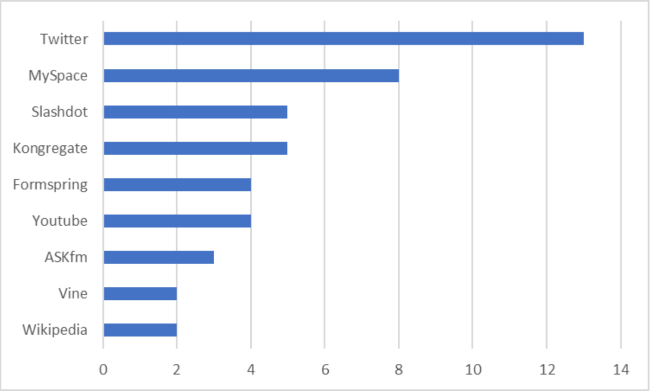

Fig. IV: Usage of various datasets

There are various electronic media for cyberbullying– such as SMS, MMS, chat rooms, forums, Email, and social networking sites (YouTube, Twitter, SnapChat, and Facebook, etc.). The major work in the literature targeted on social as main source of data because of its free access in the community zone. The SMS, emails, chatrooms and MMS are the private ways of correspondence and the communications through these e-media are more averse to be freely accessible. The graph shown in Figure IV illustrates the measureable level of the utilization of different online social media datasets for content-based cybercrime detection practices in social media. It is clear from the figure that the majority work targeted Twitter and MySpace as most common data sources in cyberbullying detection. Various studies including the work of [17,20,22,27,28,33,34] and many others used data from twitter whereas the work of [13,30,36] among others utilized MySpace. Slashdot and Kongregate are in third position with Yin et al. [11], Nahar et al. [17], Nahar et al. [19] and Dadvar et al. [34] using messages from YouTube data.

Table IV depicts the insights of various methodologies used for detecting Content-based Cybercrime in terms of main characteristics briefly described as follows:

- *Content based features:* Due to the extremely subjective characteristic of cyberbullying detection, it is possible to have diverse effects of the same text on different folks and it is a challenging task to find the effects of these during detection time. In this direction, the authors [11-37] heavily used content-based attributes like occurrence of spelling, pronouns, document length, profanity etc. The corpus and detection technique contribute to the effectiveness of these features. This study emphasized on the work utilizing content-based features that are extractable vocabulary terms of a corpus such as punctuations, profanity keywords, pronouns etc. Figure V portrays the quantifiable range of the usage of several content-based features for content-based cybercrime recognition on online social network.

  In this review, the content-based Features are grouped as profanity, pronouns, cyberbullying keywords, n-grams, Term Frequency Inverse Document Frequency (TFIDF), Bags of Words (BoW), spelling and document length. The cyberbullying texts generally incorporate insulting and abusive language. In this study, total 8 papers out of 27 contains profanity term in the text as a sign for cyberbullying.

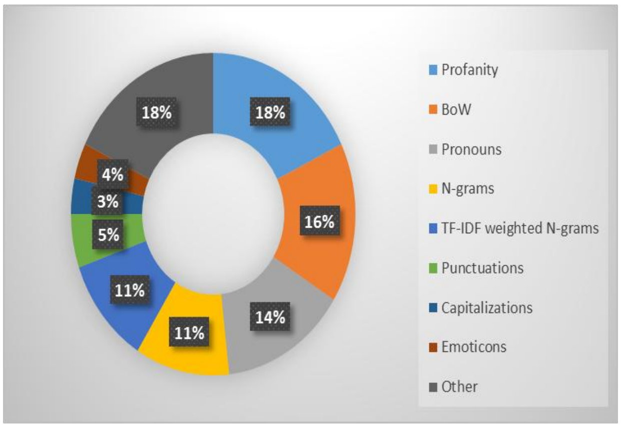

Fig. V: Exploitation (as %) of various content-based features

- *Data Pre-Processing*: It is the initial step to lessen noisy data, thus enhancing the correctness of system. It may be a two-edged weapon as the beneficial content may be vanished from the corpus during the preprocessing phase.

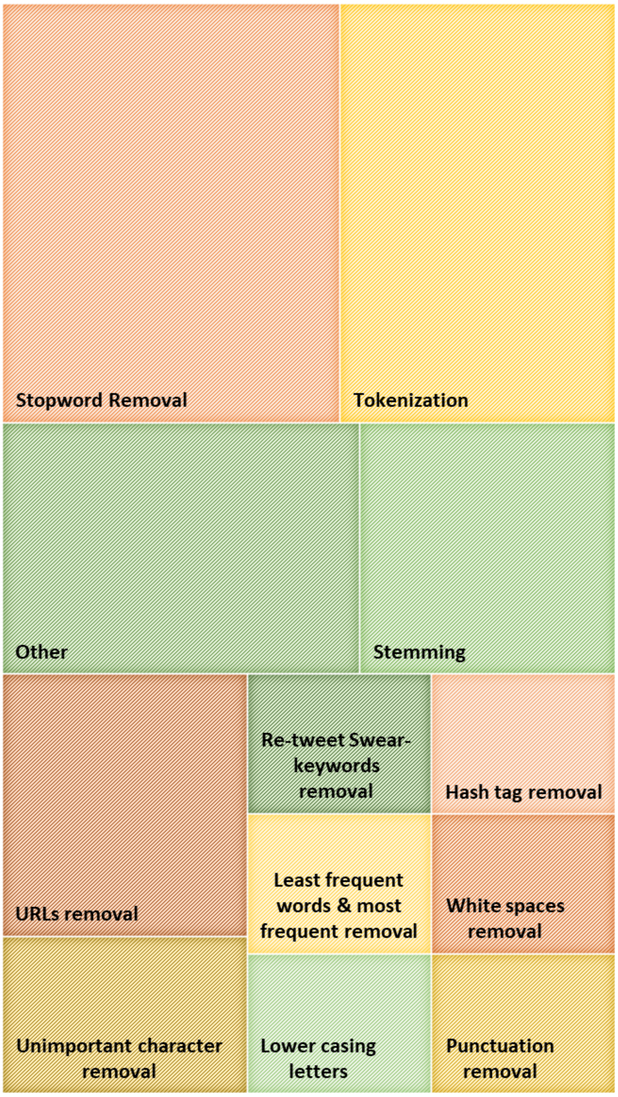

Fig. VI: Quantitative extent of the use of various pre-processing methods

TABLE. IV: Various methodologies used for detecting Content-based Cybercrime (Where NA means Not Applicable)

Literature	Preprocessing	Content-based Features Used	Dataset Used	Models/ Methodology	Evaluation Met
Yin et al. [11]	Stemming, Tokenizing, Positive instances replication	Term Frequency–Inverse Document Frequency (TF-IDF)	MySpace, Slashdot, Kongregate	SVM	Recall, Precision
Dinakar et al. [12]	Stemming, Stop words removal, Unimportant character removal	N-gram, Profanity, TF-IDF weighted unigrams	YouTube	JRip, J48, SVM, Naïve Bayes (NB)	Accuracy, Kappa
Bayzick et al. [13]	NA	Second person pronouns, Swear word, Insult word	MySpace	NA	True (Positive False (Positive/
Reynolds et al. [14]	Frequent words removal, Unimportant words removal	Bag of Words (BoW)	Formspring	Sequential Minimal Optimization (SMO), IBK, JRip, J48	Recall, Precision
Dinakar et al. [15]	Removal of stop words, Tokenizing	Profanity, BoW, TF-IDF, Weighted unigrams	Formspring, Youtube	Tree-based learner, Rule-based learner, SVM, NB	Accuracy, Kappa
Nahar et al. [16]	Removal of stop words, Stemming	BoW	MySpace, Slashdot, Kongregate	SVM	F-score, Accurac
Nahar et al. [17]	Web addresses, Re-tweet Swear-keywords, Hash tag, Stop words removal, Least frequent words & most frequent removal	TF-IDF unigrams	MySpace, Slashdot, Kongregate, Twitter	Ensemble	Recall, Precision
Dadvar et al. [18]	Stop words removal, Stemming	Profanity, Emoticons, Message length, N-gram, Bully keywords, Pronouns	Youtube	SVM	Recall, Precision
Nahar et al. [19]	NA	Special Characters, Capitalization, Profanity, Pronouns	MySpace, Slashdot, Kongregate	K-FSVM, Logistic Regression (LR), Radom Forrest (RF), NB	Average, Reca Accuracy, F-sco
Nahar et al. [20]	Lower casing letters, Stop word, Most and least frequently used words, retweet keywords, hash tag, web addresses removal	TF-IDF unigrams	MySpace, Slashdot, Kongregate, Twitter	Ensemble	Recall, Precision
Huang et al. [21]	NA	Emoticons, Punctuation, Capitalization, Profanity	Twitter	ZeroR, SMO, NB, J48,	True positive r Operating (ROC)
Mangaonkar et al. [22]	Tokenization	N-gram	Twitter	SVM, LR, NB	Recall, Precision
Hee. et al. [23]	Lemmatization, Tokenization, PoS-tagging	Character trigram BoW, Bigram BoW, Word unigram	ASKfm	Linear SVM (LSVM)	F-score
Hosseinmardi et al. [24]	Unimportant character removal, Stop words removal	Number of posts within interval less than one hour, Number of comments for the image, n-gram (n=1, 2, 3)	Instagram	LSVM	Recall, Precision

Al-garadi et al. [25]	Spelling correction, Removal of white spaces, Lowercase conversion	First and Second person, Profanity	Twitter	KNN, RF, SVM, NB	F-score, Recal Area under Curv
Galán-García et al. [26]	Tokenization	TF–IDF, N-gram	Twitter	NB, KNN, RF, J48, SMO	AUC, TPR, FPR
Singh et al. [27]	Tokenization	Part of speech tags, Question marks, Density of uppercase letters, count of exclamation points, count of smileys, Density of bad words	Twitter	Synthetic Minority Oversampling Technique (SMOTE)	Recall, Precisio F-score
Zhao et al. [28]	Tokenization	Profanity, Ensemble BoW	Twitter	LSVM	Recall, Precision
Dani et al. [29]	Stemming, Stopwords removal	Bigrams, List of profane words	MySpace, Twitter	Sentiment Informed Cyberbullying Detection (SICD)	F-score, AUC
Raisi et al. [30]	Stop word, Duplicate tweet, Retweets, emojis, URLs, punctuation removal	N-gram (n=1, 2)	Twitter, Instagram, Ask.fm	Participant Vocabulary Consistency (PVC)	Precision
Singh et al. [31]	Unimportant character, Stop words removal	Number of dashes (punctuation), Drives for reward, Tentativeness, Causation implied in comments, Sadness, Anger in comments, Positive emotions, Word count	Instagram	SMOTE	Accuracy, ROC
Zhao et al. [32]	URLS replaced by predefined characters, Tokenization	BoW, Cyberbully keywords, Profanity	MySpace, Twitter	LSVM	Accuracy, F-sco
Agrawal et al. [33]	NA	NA	Wikipedia, Twitter, Formspring	Deep learning	Recall, Precision
Dadvar et al. [34]	Punctuations, Stopwords removal	NA	Wikipedia, Twitter, Formspring, YouTube	Deep learning	Recall, Precision
Rafiq et al. [35]	Tokenization	Negative comments, Total negative words, Unigrams	Vine	AdaBoost, LR	Recall, Precision
Van Hee et al. [36]	White spaces, Abbreviations, Hyperlinks, Tokenization removal	Character n-gram BoW, Word n-gram BoW	Posts of ASKfm in Dutch and English	LSVM	AUC, Recall, F-score Accurac
Cheng et al. [37]	NA	NA	Vine, Instagram	LR, LSVM, RF	Micro F1, Macro

For example, converting uppercase text to lowercase; inadvertently may result in losing the context as capitalization is generally used to signify uproar in written communication. In the articles reviewed, 22 papers from the considered sample accomplish the pre-processing step. Here in, stemming and tokenization are most frequently used pre-processing phases. Stemming is generally accomplished on a text with BoW and n-gram TFIDF as feature set.

Through stemming, the importance of stemmed words within the datasets is highlighted by collapsing stemmed words into a term/stem. The tokenization step splits the text into an order of distinct words thereby representing a document as a set of its phrase. Figure VI Quantitative extent of the use of various pre-processing methods. The studies done in our survey include other key pre-processing tasks such as stop words removal that seem, by all means, to be of little importance to the area being referred to. In some cases, the stop word elimination can also unintentionally remove significant terms. It would be better to initially find whether the Stopwords are utilized in frequently used sentences prior to their removal.

- *Techniques for Cyberbullying Detection:* From the survey, it is revealed that majority of the work involved supervised learning techniques in the area of detection of bully content. The work of Yin et al. [11] was the original work founded during topic exploration in cyberbullying detection. Figure VII represent quantitative extent of the use of various evaluation parameters. The key evaluation parameters in supervised learning are Accuracy, Precision, Recall and F-Score. Nevertheless, the work in many papers provided experimental results using these metrics; still these studies cannot be compared directly due to the usage of diverse set of datasets. Moreover, the research works that conducted their experimental work on the same dataset, are inclined to use to extract different samples from the same dataset.

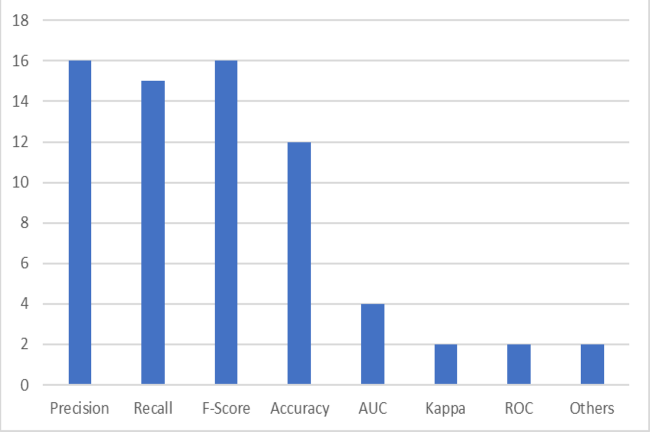

Fig. VII: Usage count of various evaluation parameters

Figure VIII depicts the application of several models/techniques organized year wise for content-based cybercrime detection in online social sites. It is clear from the figure that SVM and NB are the commonly used classification techniques for attaining effective outcomes for the detection of bully content on social networks by LSVM, RF, LR, J48, SMO, deep learning and other techniques. The graph shown in Fig. IX portrays the quantifiable range of the usage of several practices for cyberbullying detection in social media platforms.

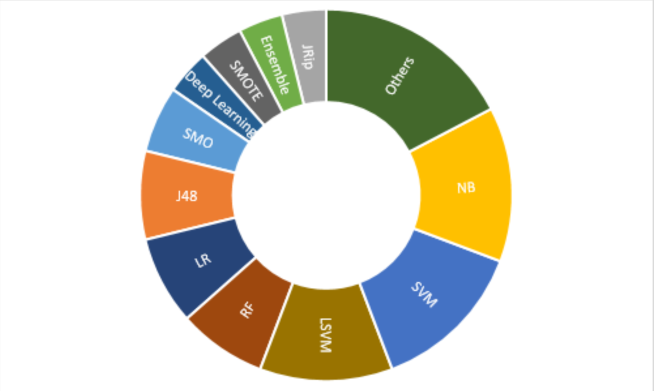

Fig. VIII: Proportion of usage of various machine learning methodology

Figure IX depicts the year-wise count of published work in the field of content-based cybercrime detection in online social media. It is clear from the figure that there is a progress in the research work in the field of cyberbullying detection from 2009 onwards and this growth is increasing at a faster pace.

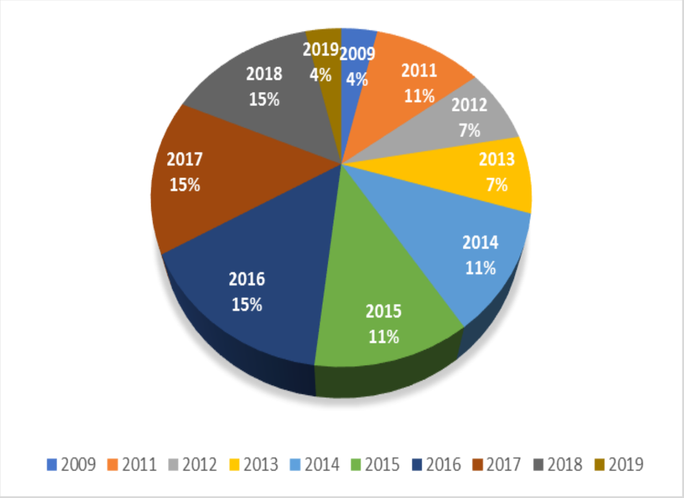

Fig. IX Year-wise cumulative assessment of published work

Integrating the outcomes of every single research, as a meta-survey, it is perceived that the usage of machine learning methods for content-based cybercrime detection delivered the intellectual methodical prototype necessary for bullying behavior and activities prediction in online social media. It is an auspicious course of study with hands-on area that predominantly depends on discovering the expandability of individual demonstrations through documented web-data available.

IV. CONCLUSIONS & FUTURE DIRECTIONS

Detection of content-based cybercrime and the facility of consequent protective actions are the key progressions of act in combating cybercrime. Although many studies and researches are dedicated to deal with this problem, there are still shortcomings that are needed to be addressed in order to wipe cybercrime out for good, eliminating it’s sad and negative consequences. After literature survey, the following research gaps have been identified:

1. From the studies, it has been observed that the research in the field of content-based cybercrime detection is

conducted from social and technical perspectives independently, neglecting the benefit of integration of other field's findings in their studies. The social studies have purely dived into causes of online misbehaviors, psychological, behavioral, personal reasons but their proposed solutions fail to incorporate the technical attributes and feasibilities of internet and social networks.

2. Most technical studies on detection of content-based cybercrime are generally static: they are unfit to deal with imbalanced, streaming, and noisy data in an efficient manner.
3. They mainly concentrate on cybercrime incidents detection after they happened, while little attention has been paid for the possibility of tools contributing to prevent the cybercrime from harming others by detecting the bullies.
4. There are insufficient proper training datasets in the field of content-based cybercrime detection. The count of harassment messages posted daily are very sparse in comparison to millions of messages that are posted in each second. Gathering suitable training data is a major challenge, as random sampling will capture only a limited bully message.

Although harassing messages are presented regular looking at on a huge number of posts messaged each second, they are exceptionally scanty. Gathering enough preparing information is a major test, since irregular examining will prompt couple of menace messages.

While most of the research papers targeted in the present study are primarily based on bullying as text, videos and images can also be used as an online system of harassment, and their effects maybe even more harmful. Another future direction from this research will be to expand the identity of the role to reflect these additional roles and also to find out whether and how people change additional roles during the bullying episode or accept it. For example, will the bystanders be persecuted or become a protector? After an incident with cyberbullying, determining the incidence and emotional status of the victims is another emerging area of research. Research is important in this area because managing the final outcomes of such situations is equally important when bullying instances are left undetected.

REFERENCES

1. [Online], http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
2. K. M. Finklea, and A. T. Catherine A. "Cybercrime: conceptual issues for congress and US law enforcement." Congressional Research Service, Library of Congress, 2015.
3. M. Thangiah, S. Basri, and S. Sulaiman, "A framework to detect cybercrime in the virtual environment," In 2012 International Conference on Computer Information Science (ICCIS) , 2012, pp. 553–557, vol. 1.
4. Q Li, "Cyberbullying in schools: A research of gender differences." School psychology international 27, no. 2, 2006, pp. 157-170.
5. R. A. Bonanno, and S. Hymel, "Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying." Journal of youth and adolescence 42, no. 5, 2013, pp. 685-697.
6. S. Hinduja, and J. W. Patchin, "Bullying, cyberbullying, and suicide." Archives of suicide research 14, no. 3, 2010, pp. 206-221.
7. J. W. Patchin, and S. Hinduja, "Bullies move beyond the schoolyard: A preliminary look at cyberbullying." Youth violence and juvenile justice 4, no. 2, 2006, pp. 148-169.
8. R. S. Griffin, and A. M. Gross, "Childhood bullying: Current empirical findings and future directions for research." Aggression and violent behavior 9, no. 4, 2004, pp. 379-400.
9. N. E. Willard, "Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress," Research Press, 2007.
10. T. Beran, and Q. Li, "The relationship between cyberbullying and school bullying," The Journal of Student Wellbeing, vol. 1, no. 2, 2008, pp. 16-33.
11. D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0." Proceedings of the Content Analysis in the WEB 2, 2009, pp. 1-7.
12. K. Dinakar, R. Reichart, and H. Lieberman. "Modeling the detection of textual cyberbullying." In fifth international AAAI conference on weblogs and social media. 2011.
13. J. Bayzick, A. Kontostathis, and L. Edwards. "Detecting the presence of cyberbullying using computer software.", 2011, pp. 93-96.
14. K. Reynolds, A. Kontostathis, and L. Edwards. "Using machine learning to detect cyberbullying." In 2011 10th International Conference on Machine learning and applications and workshops, IEEE, vol. 2, 2011, pp. 241-244.
15. K. Dinakar, B. Jones, C.Havasi, H. Lieberman, and R. Picard. "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." ACM Transactions on Interactive Intelligent Systems (TiiS) 2, no. 3, 2012, p. 18.
16. V. Nahar, S. Unankard, X. Li, and C. Pang. "Sentiment analysis for effective detection of cyber bullying." In Asia-Pacific Web Conference, Springer, Berlin, Heidelberg, 2012, pp. 767-774.
17. V. Nahar, X. Li, C. Pang, and Y. Zhang. "Cyberbullying detection based on text-stream classification." In The 11th Australasian Data Mining Conference (AusDM 2013), 2013.
18. M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. "Improving cyberbullying detection with user context." In European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2013, pp. 693-696.
19. V. Nahar, S. Al-Maskari, X. Li, and C. Pang. "Semi-supervised learning for cyberbullying detection in social networks." In Australasian Database Conference, Springer, Cham, 2014, pp. 160-171.
20. V. Nahar, X. Li, H. L. Zhang, and C. Pang. "Detecting cyberbullying in social networks using multi-agent system." Web Intelligence and Agent Systems: An International Journal 12, no. 4, 2014, pp. 375-388.
21. Q. Huang, V. K. Singh, and P. K. Atrey. "Cyber bullying detection using social and textual analysis." In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, ACM, 2014, pp. 3-6.
22. A. Mangaonkar, A. Hayrapetian, and R. Raje. "Collaborative detection of cyberbullying behavior in Twitter data." In 2015 IEEE international conference on electro/information technology (EIT). IEEE, 2015, pp. 611-616.
23. C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. "Automatic detection and prevention of cyberbullying." In International Conference on Human and Social Analytics (HUSO 2015), IARIA, 2015, pp. 13-18.
24. 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." Computers in Human Behavior 63, 2016, pp. 433-443.
25. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. "Detection of cyberbullying incidents on the instagram social network." arXiv preprint arXiv:1503.03909, 2015.
26. P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas. "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying." Logic Journal of the IGPL 24, no. 1, 2016, pp. 42-53.
27. V. K. Singh, Q. Huang, and P. K. Atrey. "Cyberbullying detection using probabilistic socio-textual information fusion." In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE Press, 2016, pp. 884-887.
28. R. Zhao, A. Zhou, and K. Mao. "Automatic detection of cyberbullying on social networks based on bullying features." In Proceedings of the 17th international conference on distributed computing and networking, ACM, 2016, p. 43.
29. H. Dani, J. Li, and H. Liu. "Sentiment informed cyberbullying detection in social media." In Joint European Conference on Machine Learning and Knowledge Discovery in

Databases, Springer, Cham, 2017, pp. 52-67.

30. E. Raisi, and B. Huang. "Cyberbullying detection with weakly supervised machine learning." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, 2017, pp. 409-416.
31. V. K. Singh, S. Ghosh, and C. Jose. "Toward multimodal cyberbullying detection." In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, 2017, pp. 2090-2099.
32. R. Zhao, and K. Mao. "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder." IEEE Transactions on Affective Computing 8, no. 3, 2016, pp. 328-339.
33. S. Agrawal, and A. Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms." In European Conference on Information Retrieval, Springer, Cham, 2018, pp. 141-153.
34. M. Dadvar, and K. Eckert. "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study." arXiv preprint arXiv:1812.08046, 2018.
35. R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra. "Scalable and timely detection of cyberbullying in online social networks." In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, ACM, 2018, pp. 1738-1747.
36. C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. "Automatic detection of cyberbullying in social media text." PloS one 13, no. 10 (2018): e0203794.
37. L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu. "XBully: Cyberbullying Detection within a Multi-Modal Context." In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM, 2019, pp. 339-347.

AUTHORS PROFILE

Amanpreet Singh is currently doing PhD in field of content-based cybercrime detection in Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology, Patiala. He received his Bachelor’s and Master’s degree from Thapar Institute of Engineering & Technology. He has done research in his masters in the field of VLSI physical design automation using Evolutionary approach. His major research interests include Meta-heuristic algorithms, cybercrime analysis and Swarm intelligence. His current research includes the application of machine learning and swarm intelligence techniques for content-based cybercrime detection.

Maninder Kaur is holding an academic position as Assistant Professor in Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology, Patiala. She received her Bachelor Degree from Sant Longowal Institute of Engineering and Technology and Master’s degree from Punjabi University. She completed her Ph.D. in the field of VLSI physical design automation using Evolutionary approach. Her major research experiences and interests include IOT, Big Data Analytics, Data Mining and Swarm intelligence. She has 25 publications including prominent journals, conferences. She has also acted as mentor in various capstone projects in the field of IOT and machine learning. She is currently supervising two Ph.D. and two M.E. students. She has also supervised more than 11 M.E./M.Tech. thesis. She is an associate member of The Institution of Engineers, India. Her current research includes the application of machine learning and swarm intelligence techniques for big data analytics, cyberbullying and sarcasm detection.