Cyber Bullying and Machine Learning: A Survey

Ibtihaj Alanazi
Center for Secure and Dependable Systems
University of Idaho
Moscow, ID 83844, USA
alan3060@vandals.uidaho.edu

Jim Alves-Foss
Center for Secure and Dependable Systems
University of Idaho
Moscow, ID 83844, USA
jimaf@uidaho.edu

***Abstract*—On an average, 20% to 40% of all teenagers have been mistreated online. With appropriate detection of possible harmful messages, successful prevention can be achieved. However, there is a requirement for intelligent systems to identify possible risks automatically, given how the Web is overloaded with massive information. This is what encouraged us to control bullying by detecting it on different social media sites so that the people out there can take initiatives to end it. This paper surveys several attempts to use machine learning technology to detect instances of cyber bullying and summarizes their results. Overall, the results are promising, but with these results still have room for improvement**

Keywords-component; cyber bullying; machine learning

I. INTRODUCTION

The rapid development of online communication and information sharing platforms and the enthusiastic participation of their users have enabled peer-to-peer communication at unprecedented scale and diversity. On one hand, these communication channels, such as online social networks and news sharing websites, o□er myriad opportunities for knowledge sharing and opinion mobilization. On the other hand, they also serve as a fertile domain for an abundance of unfortunate intimidation and hateful aggression and cyberbullying towards individuals targeted because of their identities or expressed opinions. Cyberbullying can also pose individual health costs ranging from anxiety and depression to severe outcomes such as suicide. A study by the Pew Research Center [1] found that 60% of the US Internet users have experienced cyberbullying online, with young women enduring particularly severe forms of it. According to the Pew Research Center [2] around 62% people think cyberbullying is a major problem.

Engaging with various online social media platforms multiple times each day is common for citizens of today across most demographics. An ever-increasing number of individuals share their opinions, personal experiences, and social views through various social media platforms. Social media sites with significant user base include social networking sites such as Facebook, Twitter and ASKfm, photo and video sharing sites such as Flickr and YouTube, social news sites such as Reddit and Digg, blogging platforms such as WordPress and Blogger, and social messaging platforms such as WhatsApp and Snapchat. Because of the ubiquitous access to the Internet and availability of wireless personal communication devices, users can now readily engage with social media platforms and express their views and opinions on diverse topics at any time and from almost any location at their convenience.

Millions of posts, in the form of texts, images, and videos are appearing daily on popular social media platforms. Authors of those posts write about their life, share opinions on a variety of topics and discuss current issues. As more and more users engage with these sites, they become valuable repositories of people's opinions and sentiments about the services they use as well as their political and religious views. Hence, these sites have key influence on user's opinions and sentiments. Correspondingly, the collected data serve as valuable data sources for businesses, researchers, and policymakers. Whereas these new communication channels, such as online social networks [3] and news sharing sites [4], o□er myriad opportunities for knowledge sharing and opinion mobilization [5], they also reveal an abundance of unfortunate fear and aggression [6] towards individuals targeted because of their expressed opinions or identities. This nasty and often coordinated victimization of individuals have significant social costs ranging from social ostracism to opinion marginalization and suppression and can cause severe individual health detriments such as anxiety [7] to depression [8] to suicide ideation [9].

The National Crime Prevention Council reported in 2011 that cyberbullying is a problem that affects almost half of all American teens. The consequences of cyberbullying are similar to traditional bullying, and have been shown to include depression, low self-esteem and suicide attempts [35],[36]. However, in some cases the consequences of cyberbullying can be more severe and longer lasting due to some specific characteristics of cyberbullying. Cyberbullying can be undertaken 24 hours a day, every day of the week, and unlike traditional bullying, it is independent of place and location [37]. Moreover, online bullies can stay anonymous [38] and being bullied by an unknown person can be more distressing than being bullied by someone familiar (Kowalski et al., 2012). Furthermore, anonymity triggers cyberbullying behavior for people that would not bully face-to-face [39].

Online materials spread very fast and in a couple of minutes thousands of Internet users can have access to it [38]. There is also the persistency and durability of online materials and the power of the written word [39]. In the case of cyberbullying through text, the targeted victim and bystanders can read what the bully has said over and over again, and also in the case of images the hurtful content can stay online for a long period of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

time and if tagged with the name or other personal features of the victim it will keep showing.

A. Sentiment Analysis

Sentiment analysis is the study of computationally detecting and categorizing sentiments expressed in a piece of text or in a whole content, especially in order to decide whether the writer's attitude towards a certain topic is positive, negative, or neutral. It is a combination of natural language processing, text analysis and computational linguistics. In this process a sentence is considered positive if it has positive keywords and is considered negative if it has negative keyword. The comparison among the number of each type of content decides the positivity and negativity of the whole content. This study tends to provide an algorithm that may help in analysis of words that may lead to crime detection especially in social sites.

B. Motivation

Bullying can cause depression and sometimes even suicides by the victim. It effects the victim both mentally and physically. It has adverse impacts not only on the victims but also on those who bully and those who witness bullying. Consequently, it increases crimes, mental and physical illness and causes the victims to isolate themselves. As a response to cyber threats, several national and cross-national child protective initiatives (e.g., Suicide Prevention Resource Center (https://www.sprc.org/), Stop Bullying (https://www.stopbullying.gov/)) have been starting projects over the last few years to increase online child safety. Despite these efforts, much undesirable or even hurtful content remains online.

On an average, 20% to 40% of all teenagers have been mistreated online, as suggested by recent research reports [10]. With appropriate detection of possible harmful messages, successful prevention can be achieved. However, there is a requirement for intelligent systems to identify possible risks automatically, given how the Web is overloaded with massive information. This is what encouraged us to control bullying by detecting it on different social media sites so that the people out there can take initiatives to end it.

C. Problem Statement and Research Gap

Detecting cyberbullying is a challenging task. Several issues must be resolved with respect to the dataset, algorithm, building a better model and accuracy of result etc. Referring to the current research in the field of detection of cyberbullying there exists a gap between a high false alarm and low accuracy. This gap can be overcome with the use of optimum feature (attributes) selection, using the most appropriate machine learning algorithm and designing a classifier which will result in better detection of cyberbullying contents.

II. Literature Review

Related works in the area of cyberbullying can be partitioned into four sections. They are: definition of cyberbullying, cyberbullying research in online social networks, cyberbullying detection techniques and systems, applications, and tools for detection of cyberbullying in online social networks. Past works in each of these four areas are explored in the following four sections.

A. Definition of Cyberbullying

Cyberbullying is defined as an aggressive, intentional act that is carried out by a group or an individual, using electronic, digital, multi-modal forms of contact, messaging, communication, repeatedly against a victim who cannot easily defend him or herself [11]. One huge distinction between traditional bullying and cyberbullying is that the perpetrator of cyberbullying really wants to hurt the feelings of the victim [12]. Intending to hurt the feelings of the victim, imbalance of power and the repetitive nature of contact are the unique traits of cyberbullying. Although cyberbullying is sometimes defined as an electronic form of face-to-face bullying rather than a distinct phenomenon [13], considering cyberbullying as merely the electronic form of face-to-face bullying may overlook intricacies of these behaviors, such as repetition of aggression and imbalance of power in an electronic context. Repetition in cyberbullying is problematic to contextualize, as there can be di□erences between the perpetrator and victim when it comes to the conceptualization of how many incidents occur and their potential consequences. A single aggressive act such as uploading an embarrassing picture to the internet can result in continued and widespread ridicule and humiliation for the victim. While the aggressive act is not repeated, the damages caused by the act is re-lived by the victim through an elongated humiliation. Power imbalance in an electronic context can be defined as the perpetrators having superior technological skills or the victim being “shy” or “modest” and the perpetrator knowing the victim in real world [12]. From over-viewing the existing literature, around eight types of cyberbullying behaviors can be recognized [14]:

1. *Flooding* involves the bullies sending repeated frequent nonsensical comments/posts in order to not allow the targeted victim to participate in the conversation
2. *Masquerade* involves the bullies pretending to mimic or impersonate the target victim
3. *Flaming/Bashing* involves an online fight where the bully sends and/or posts insulting, hurtful and vulgar contents to the targeted victim privately or publicly in an online group
4. *Trolling* involves purposely publishing comments which disagree with other comments in order to incite arguments or negative emotions although the comments themselves might not be vulgar or hurtful in themselves
5. *Harassment* is the kind of conversation where the bullies frequently send insulting and rude messages to the victim privately.
6. *Denigration* occurs when the bullies send or publish gossips or untrue statements about the victims to damage the victims’ friendships/reputations

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

7. *Outing* occurs when bullies send or publish private or embarrassing information in public chat-rooms or forums. This type of cyberbullying is similar to denigration. However, in the outing, there might be a relationship between bully and victim.
8. *Exclusion* involves intentionally excluding someone from an online group. This type of cyberbullying happens among youth and teenagers more prominently.

B. Cyberbullying Research on Online Social Networks

Analysis and detection of cyberbullying/profanity/harassing incidents in several online social networks like Twitter , Ask.fm, YouTube, FormSpring, chat-services have been performed by different groups of researchers.

Twitter is a text-based social network where a user can update their status by not more than 280 characters. Opinion mining and sentiment analysis techniques have been used to detect cyberbullying in Twitter. Al-garadi et al. [15] used a negative word list to streamline the tweets that contained those negative words . After that, a sentiment classifier was built with four classes: negative with bullying intentions, negative without bullying intentions, positive or good content and neutral. The labeling of the tweets was performed using the Amazon Mechanical Turk platform which was then employed to build and evaluate the classifier. The reported results were 67.3%.

The relationship between cyberbullying and anonymity in online social networks has been explored in depth as well by Hosseinmardi et al. [16]. Ask.fm is a semi-anonymous online social network, where the users have the option to hide their identity when posting questions/comments on a profile. The work by Hosseinmardi et al. [16] used snowball sampling [17] and collected 30,000 user profiles. These profiles were then analyzed using interaction graphs, word graphs, frequency distributions and network properties such as reciprocity, clustering coe□cient, and the influence of negativity on in-degree and out-degree. It was found that the most vulnerable users were the least active in terms of online social network activity, such as receiving/posting likes.

Research based on tracking and categorization of internet predators on online chat services has also been performed by Kontostathis et al. [18]. 288 chat-logs were collected from perverted-justice.com, a project where the volunteers pose as teens and tweens to trap potential sexual predators. Identified categories of the terms and phrases frequently used by the predators were: deceptive trust development, grooming, isolation, and approach. The idea was to distinguish between predators and victims and to this aim, their developed clustering methods were able to achieve an accuracy of 93%. This experiment used 29 transcripts.

Research has been performed to detect instances of harassment in online social networks and chat services as well. Yin et al. [19] partition online social networks into two groups: discussion style and chat style. In discussion style environments, there are various threads, usually with multiple posts that populate each of those threads. Users can start a new thread or participate in an existing thread by posting comments. Each thread contains posts that adhere to a predefined topic. On the other hand, in chat style environments, ongoing conversations are more casual and usually, each conversation only consists of a few words with little information. Topical and sentimental features were used to train the supervised classifier to detect harassment after collecting data from Kongregate (chat style) and MySpace (discussion style). The performance of N-grams was lower than the TFIDF weights features. The Precision of 0.394 and F-score of 0.481 was the highest.

It will be interesting to have further insights into cyberbullying behavior in multi-modal online social networks like Vine and Instagram where users can share videos and images, respectively. In comparison to textual cyberbullying, these social networks also provide potential perpetrators with a platform on which to harass the victim though posting harmful images or insulting videos instead of just posting mean comments. Moreover, an in-depth analysis of the correlation between the media contents and cyberbullying behavior can also better our understanding of cyberbullying behavior in online social networks. Finally, delving into the details of cyber-aggression and cyberbullying and investigating the potential distinguishing factors between these two behaviors are also some untapped areas of future research.

C. Cyberbullying Detection Techniques

This section briefly outlines research focusing on e□ective and e□cient cyberbullying detection techniques.

Manual analysis of data and establishment of relationships between multiple data items are often prone to errors. Machine learning can address such challenges and can be successfully applied to these problems. To apply machine learning algorithms, an input dataset is created comprising of instances described by a set of features. These features can be continuous, categorical or binary. When the data instances are associated with known labels, the learning is termed supervised machine learning [20]. In contrast, in unsupervised machine learning [21], data instances are unlabeled. Unsupervised algorithms are applied to datasets to discover unknown, but potentially useful classes or groups of items. The learner is not provided with any direct guidance about which actions to take but must discover which actions yield the best result, by systematically exploring available options. Supervised machine learning algorithms are used to monitor whether instances of data are classified correctly, misclassified or assigned relatively high likelihoods of belonging to the particular category. Unsupervised Machine Learning algorithms are used to analyze how data can be grouped into clusters and inter-cluster relations. This section gives a brief review of machine learning techniques employed by previous studies for detection of cyberbullying.

Research has been proposed based on the text mining paradigm for detection issues that are closely related to cyberbullying such as such as online sexual predator recognition [22] and spam detection [28]. Modeling the detection of textual cyberbullying has been a cornerstone of cyberbullying research [31] where the problem of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

cyberbullying detection in Twitter was decomposed into a problem of detecting discussions on sensitive topics, thus rendering the problem into a text classification sub-problem. Three topics were identified as sensitive: sexuality, race/culture, and intelligence. Upon collecting comments pertaining to the aforementioned sensitive topics, the final step was to determine the profanity content of those comments in order to detect cyberbullying. JRip classifier [33] was reported to be the best performing classifier in this technique with a F1-score of 78.

Comparison of di erent approaches to building e ective machine learning classifiers for cyberbullying have also been investigated [27], namely human expert system, supervised machine learning models and a hybrid system combining both machine learning and expert systems. Labeled data from YouTube was used to evaluate each of these three systems. In the evaluation, it was reported that the expert model outperformed all of the machine learning models. The machine learning models' sensitivity to the class skew of the data-set (10% bullying and 90% non-bullying) was attributed to this under-performance. The hybrid approach was reported to have performed better than both the expert model and the machine learning model. Other techniques such as building query terms of phrases and words pertaining to cyberbullying have been developed in the past to detect instances of cyberbullying. Kowalski et al. [13] used labeled data from FormSpring.me and went on to build the most e ective query terms for e cient detection of cyberbullying leveraging two models: language and machine learning. It was reported that the terms generated by the machine learning model were the better performing one, yielding both high recall and precision than its language model counterpart.

Initial work in cyberbullying detection techniques has mostly concentrated on the conversations' content though they did not attend to the characteristics of the actors involved in cyberbullying. Social studies demonstrated that men and women bully each other in a di erent way. For example, women tend to employ aggressive communication styles, such as excluding someone from a group of conspiracy against them whereas men tend to use more words and phrases threatening outrage. Lee and Ma [4] reported that pronouns like "I", "you", "she", etc. are used more by females and noun specifiers such as, "a", "the", "that" are used prominently by males. These findings motivated several cyberbullying researchers to include gender-specific information in cyberbullying detection techniques. Gender-specific information in online social networks has been reported to be useful in improving the performance of a cyberbullying detection system [26] with an out-degree centrality scores 0.571 vs 0.33.

Graph models in social networking sites have also been actively used in cyberbullying research. Hosseinmardi et al. [16] presented a graph model to extract a cyberbullying network. This then led to identifying the most active predators and victims through a ranking algorithm. They improved the classification performance by applying a weighted TF-IDF function, in which bullying-like features were scaled by a factor of two. Techniques to detect cyberbullies and cyber-predators have also been proposed in the past [28]. A cyber predator is a person who uses the Internet to hunt for victims to take advantage of them in several ways, including sexually, emotionally, psychologically or financially. Cyber predators know how to manipulate kids, creating trust and friendship where none should exist [11]. Online sexual predator related research identified communication and text-mining techniques to di erentiate predators and victims by analyzing the one-to-one conversations [23]. Rezvan et al. [28] partitioned the online predator detection problem into two sub-problems, namely identifying predators and recognizing predator's conversation techniques/lines for identifying them. Three stages were then proposed: pre-filtering stage, feature extraction stage, and classification stage. For the feature extraction stage, two categories of features were leveraged: lexical and behavioral features [15]. Lexical features were described as those features that could be derived from the raw text of the conversation between the victim and the potential predator, for example, unigrams and bigrams [13], number of emoticons used and the weighted TF-IDF or the cosine similarities. The behavioral features included the number of questions asked, intention (grooming, hooking) to capture the action of the users [15]. For classifying predators, several approaches were investigated by the researchers, namely, decision trees [16], Neural Network [13] and Maximum-Entropy [31].

Andriansyah et al. [22] conducted research on the classification of cyberbullying comments on Instagram using Support Vector Machine (SVM). For a dataset, they choose the comments from accounts of Indonesian celebrities namely, Karin Novilda and Samuel Alexandar. A total of 1053 comments were taken as a training dataset and 34 as a test dataset. For implementing the method, firstly, they created a text term matrix with R language to develop SVM model. Once the development of SVM mode is completed, they used it to predict whether a comment is cyberbullying or not. They achieved an accuracy of 79.41 %.

Eshan and Hasan [23] worked on the application of machine learning to detect abusive Bangla texts. They consider various machine learning algorithms and compare which one is better. Their experiments include algorithms such as Multinomial Naïve Bayes (MNB), Random Forest (RF) and Support Vector Machine. For the preparation of dataset, they collected data from the account of Bangladeshi Facebook celebrities. Only Bengali Unicode was considered, and all other special characters like @, - etc. were removed. For the validation of results, they use 10 folds cross-validation method. Using this method, they were able to detect 50% of the abusive words. Furthermore, the experiments were conducted with three types of string features: unigram, bigram and trigram. After that, unigram, bigram, trigram features are extracted from all of the comments and vectorized using CountVectorizer and TfidfVectorizer. After vectorization, their results show that in all cases SVM with a linear kernel shows the highest accuracy level. Lastly, they concluded that trigram TF-IDF Vectorizer features with SVM linear kernel gives the higher accuracy of 82% among all the algorithms.

Noviantho et al. [24] constructed a classification method using SVM with several kernels and Naïve Bayes. They compared their methodology with the research of Reynolds et al. (2011) who used decision tree and k-NN. The data used to create a dataset includes conversation messages taken from the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Kaggle (www.kaggle.com). They proceeded with data preprocessing, extraction, classification and lastly evaluation. They divided the data into 2, 4 and 11 classes. After completing text extraction, they classified through Naïve Bayes, SVM with linear, Poly, RBF and sigmoid kernels. Then they evaluated the accuracy rate with the method of the confusion matrix. Based on their model, SVM gave the best average result of 91.95 % and SVM-RBF gave the worst average result of 86.73 % for 11 classes. On the basis of n-grams, the best average result attained by n-gram was 92.75% and the worst one was 89.05%.

Nurrahmi and Nurjanah [25] detected cyberbullying using SVM. For dataset, they used twitter posts. Those posts were harvested from twitter by using a web scraper tool Selenium. Selenium used Chrome driver and open the URL for doing queries for twitter login, then requested data in the form of the HTML format and parsed it to get the required data. To harvest all the data, they used a step called scroll event before parsing. After that, they preprocessed the harvested data. This step includes removing special characters, URLS, identical twitter with similar text content and images and symbols from posts. They obtained 301 cyberbullying tweets, 399 non-cyberbullying tweets, 2,053 negative words and 129 swear words. They used the SVM and K-nn for the classification of cyberbullying. SVM achieved the highest F1-score of 67%.

Huang et al. [26] worked on cyberbullying detection using social text analysis to improve the accuracy of cyberbullying detection. In this research, they used the corpus data set and apply Synthetic Minority Oversampling (SMOT) Technique, in which they apply six algorithms like bagging, j48, SMO, Dagging, NaïveBayes and ZeroR and compare all the results. Dagging gave the highest RoC of 0.755.

Ozel et al. [27] conducted the first study to detect cyberbullying from Turkish texts. They created a dataset from Instagram and twitter messages and applied machine learning techniques such as: Support Vector Machine (SVM), Decision Tree, Naive Bayes Multinomial (MNB) and k-Nearest Neighbor (kNN) to detect and classify cyberbullying. The dataset was constructed manually consisting of 900 twitter and Instagram messages. Half of the messages (450 messages) were cyberbullying content and another half was cyberbullying unrelated content. Half of the cyberbullying contents (225 messages) were written by male users and the other half were written by female users. Two well-known feature selection methods Chi-Square and Information Gain were applied to show whether feature selection improves the classification accuracy or not. Then they applied Decision Tree, Naïve Bayes Multinomial, SVM(Support Vector Machine) and k- Nearest Neighbor classifiers to each fold for both datasets, calculated the F-measure values and took the average of the F-measure values for five folds. This gives a baseline result. There were two types of datasets, one with emoticons and another without emoticons. In comparison the dataset with emoticons had the better classification accuracy. The feature selection method Chi-Square and Information Gain both gave quite similar results, but Information Gain had slightly better accuracy. In terms of accuracy, Naïve Bayes performed the best when features were not applied and k- Nearest Neighbor was the most accurate when features were applied. The accuracy of all classifiers improves except for Decision Tree when feature selection is applied. The accuracy of SVM was lower than Naïve Bayes and k- Nearest Neighbor in most of the cases because the parameters were not optimized. In terms of running time, Naïve Bayes became the best classifier in terms of both training and testing time with 0.37 seconds, SVM was second best with 0.75 seconds.

Rezvan et al. [28] worked on prediction of cyberbullying occurrence in media-based social media, therefore, they predicted cyberbullying from an image typically with a text caption also the comments followed by the image in America based social media accounts. They chose Instagram as their social media. For data set, they used 25,000 public account in Instagram. They collected the user's profile data which includes image with caption and comments of other users. For labeling, they used a dictionary with profane words. To design and train classifier fivefold cross validation method was applied. Also, a logistic regression was applied to train the predictor. 98% of cyberbullying activities were captured for set 0. This showed that cyberbullying incidents can be predicted with 0.99 recall for Set 0. The best false positive rate over Set 0 is 3%, using only the image contents, media and user metadata based on a ridge regression classifier.

Haidar et al. [29] worked on cyberbullying detection in the Arabic language. They have shown how NLP and some machine learning algorithm work to detect cyberbullying. The machine learning algorithms include Naïve Bayes, K-NN, SVM, Decision Tree etc. They proposed a multilingual cyberbullying detection system in Arabic language on Facebook and Twitter. Then they intended to collect dataset from Facebook and Twitter and classifying data with ML algorithms. For the performance measurement, they proposed re-call, precision, and F-measure to reach a system with optimum performance. They also did not implement any methodology to detect cyberbullying. They only proposed to apply the above methods.

Del Vigna et al. [30] developed a hate speech classifier for the Italian Language. They built a corpus of comments from Facebook public pages of Italian newspapers, politicians, artists, groups etc. They collected 17,567 comments from 99 posts from these pages. Some of the comments were annotated to one of the three levels of hate: no hate, weak hate and strong hate. The rest of the comments were annotated to one of the two levels of hate: hate and no hate. They tested these datasets with two classifiers: SVM and Recurrent Neural Network named Long Short-Term Memory (LSTM). They followed 10-fold cross validation process for each dataset. On the three-class dataset, SVM and LSTM gave 64.61% and 60.50% of accuracy respectively. On the two-class dataset, SVM and LSTM attained the accuracy of 80.60% and 79.81% respectively. We see that they produced a better result with SVM classifier. But the results of three-class dataset were not satisfactory using any of the classifiers.

Zhao et al.'s [31] work is about research on cyberbullying detection in Twitter. They approached with a whole new method called embedding-enhanced Bag of Words model (EBoW) For a dataset, they used texts or posts from Twitter. For implementing EBoW they first defined a list of insulting

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

words based on expert knowledge and linguistic resources, furthermore, they extended the insulting words to define bullying features. Different weights were assigned to bullying features based on the cosine similarity between word, EBoW. After that, based on weight they classified the intensity of cyberbullying. They took 1,762 sample post from Twitter and they got 684 post as bullying instances. They trained and tested with 5-fold comparing with BoW, sBoW, LDA, LSA. The result of EBoW came out best of all. Precision was 76.8%, Recall 79.4% and F1 Score 78.0%.

Mangaonkar et al. [32] improved the detection of cyberbully detection using collaborative computing. Their result indicates an improvement in time and accuracy of the detection mechanism over standalone paradigm. They created two datasets, and both consisted of tweets from Twitter. A balanced dataset using 170 bullying and equal non bullying contents. Another was unbalanced using 177 bullying and 1163 non bullying contents. Then they applied Naïve Bayes, SVM and Logistic Regression machine learning techniques with word tokenizer and bigram tokenizer parameter settings. With a balanced dataset, Logistics Regression performed little better than others, with more than 60% precision recall, and accuracy. Naïve Bayes was close to Logistic regression and SVM had better recall but bad accuracy and precision. With unbalanced data, Logistics Regression again performed with more than 30% correct predictions on average whereas in Naïve Bayes the values had dropped and SVM failed. After that collaboration methods namely, AND parallelism, OR parallelism and Random 2 Or parallelism were used to determine if there is any improvement on precision, accuracy and recall. Among the techniques used AND parallelism had the best accuracy and OR parallelism had the best recall and 7 out of 15 cases using collaboration techniques worked better than their sequential counterpart. This paper gave some new insights how to improve the result using collaboration techniques after using the machine learning techniques to detect cyberbullying. But they mentioned that the results achieved were without any tuning to the algorithms used so if the algorithms were a little edited maybe the result would have been much better. One interesting future work of theirs mentioned was that the history of two twitter accounts were not considered which obviously plays a vital role in detecting cyberbullying. This is one of our concerns also. SVM classifier performed poorly in this research but in most other papers SVM was defined as the best, so if SVM was tuned and used in kernel it might have performed much better.

Gorro et al. [34] aimed to detect cyberbullying actors in twitter based on texts and the credibility analysis of user and also notify them about the harm of cyberbullying. They collected the dataset from twitter. The labelled the data by building a web-based labelling tool. Their data labelling system includes registration of participants, adding negative words and swear words, calculating labeling score and updating corpus and finally labelled tweet as negative word corpus and swear word corpus. After that they preprocessed the data by tokenizing, removing symbols, number etc. Then they extracted the features and the result of this step is formed as a table. Finally, they trained the data to develop SVM and KNN. After detecting cyberbullying by these two models they found that SVM with RBF kernel (c=4) results in the highest f1-score, 67%. SVM with linear kernel and KNN is less than that of RBF kernel. During feature extraction, they measured the credibility of users and found 257 normal users, 45 harmful bullying users, 53 bullying actors and 6 prospective bullying actors.

D. Performance Measures

Evaluating the performance of cyberbully detection system is a critical process. There are several existing metrics that measure performance. The most basic and commonly used method is confusion matrix. The confusion metrics is a specific table layout that allows visualization of the performance of an algorithm, (see Table 1) In balanced datasets (ones where there are similar percentages of bullying versus non-bullying), accuracy is an acceptable metric. However, in an unbalanced dataset, accuracy is not a good metric. For example, if 5% of posts are bullying, a detection algorithm that said all posts are non-bullying, would still be 95% accurate. This result is much better than most of the reported results found in this survey. An F1 score, or separate precision and recall results are more commonly reported values for unbalanced datasets. In some of the surveyed studies it is unclear when authors reported "accuracy"

The most basic and commonly used metrics are False Positive Rate (FPR), False Negative Rate (FNR), True Positive Rate (TPR), True Negative Rate (TNR), Recall (Detection Rate), Precision, Accuracy, and F1-Score. These performance metrics are calculated from False Positive (FP), False Negative (FN), True Positive (TP), True Negative (TN) as shown in Table 1. In balanced datasets (ones where there are similar percentages of bullying versus non-bullying), accuracy is an acceptable metric. However, in an unbalanced dataset, accuracy is not a good metric. For example, if 5% of posts are bullying, a detection algorithm that said all posts are non-bullying, would still be 95% accurate. This result is much better than most of

Table 1. Confusion Matrix

	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	True Positive (TP)	False Negative (FN)
Actual Non-Bullying	False Positive (FP)	True Negative (TN)

True Negative (TP): Correctly classifying bully as bully.
True Positive (TN): Correctly classifying a non-bully as a non-bully.
False Positive (FN): Incorrectly classifying bully data as a non-bully.
False Negative (FP): Incorrectly classifying a non-bully as bully.

$$\text{False Positive Rate (FPR)} = \frac{FP}{TN+FP} \qquad \text{Precision} = \frac{TP}{TP+FP}$$

$$\text{False Negative Rate (FPR)} = \frac{FN}{TP+FN} \qquad \text{Recall(Detection Rate)} = \frac{TP}{TP+FN}$$

$$\text{True Positive Rate (FPR)} = \frac{TP}{TP+FN} \qquad \text{F1 - Score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

$$\text{True Negative Rate (FPR)} = \frac{TN}{TP+FN} \qquad \text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I. TABLE 2. SUMMARY OF REVIEWED PAPERS

Paper	Dataset	Methods	Results	Limitation
Andriansyah et al. 2017	Comments from accounts of Indonesian Selebgrams	SVM	79.41% accuracy	Use of kernels might give better result
Eshan and Hasan 2017	Comments from accounts of Banagladeshi Facebook celebrities	MNB, RF, SVM	SVM linear kernel with trigram TF-IDF 82% accuracy	Spelling checker was not implemented
Noviantho et al. 2017	Kaggle.com	SVM, Naïve Bayes	SVM with poly Kernel average accuracy 97.11%	Shortened words and spell check was not handled
Nurrahmi and Nurjanah 2018	Twitter	SVM, K-NN	SVM with highest F1-score of 67%	Stemming and spelling check wasn't done. Male and female partitioned dataset was not of any use
Huang et al. 2014	Twitter	Bagging,J48,SMO, Dagging,Naive Bayes,ZeroR	RoC of 0.755	No certain classifier the mentioned
Ozel et al. 2017	Turkish texts from Twitter and Instagram	SVM, NB, Decision Tree, K-NN	F1-score 0.81 for NVB is highest	No implementing
Mangaonkar et al. 2015	Tweets from Twitter	Naïve Bayes, SVM and Logistic Regression	Logistic Regression has above 60% precision recall, and accuracy	The result of three-class dataset is not satisfactory
Zhao et al. 2016	Tweets from Twitter	SVM with Embedding's Bag of Words (EBoW)	EboW Precision76.8%, Recall 79.4%, F1 score 78.0%	Didn't classify the dataset
Del Vigna et al. 2017	Facebook Posts	Selenium scrapper tool, SVM	SVM for two-class (80.60%) and three-class (64.61%)	Didn't try any other models which might give better result
Gorro et al. 2018	Facebook Posts	SVM	precision 88%, recall 87%	Dataset was too small. Larger dataset might be added

the reported results found in this survey. An F1 score, or separate precision and recall results are more commonly reported values for unbalanced datasets. In some of the surveyed studies it is unclear when authors reported "accuracy" if they meant the mathematical accuracy, or if they were using the term incorrectly. With this in mind, we summarize the results of our review in Table 2.2 with an understanding that some of the reported results are not directly comparable to each other or need further evaluation, which is beyond the scope of this paper.

III. REFERENCES

[1] [1] A. Geiger, "How and why we studied teens and cyberbullying," *Pew Research Center*, 2018. [Online]. Available: http://www.pewresearch.org/fact-tank/2018/09/27/qa-how-and-why-we-studied-teens-and-cyberbullying/.

[2] [2] A. Perrin, "Social Media Usage: 2005-2015," *Pew Research Center*, 2015. [Online]. Available: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/.

[3] [3] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[4] [4] C. S. Lee and L. Ma, "News sharing in social media: The effect of gratifications and prior experience," *Comput. Human Behav.*, vol. 28, no. 2, pp. 331–339, Mar. 2012.

[5] [5] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in *2010 43rd Hawaii International Conference on System Sciences*, 2010, pp. 1–10.

[6] [6] G. S. O'Keeffe, K. Clarke-Pearson, and C. on C. and Council on Communications and Media, "The impact of social media on children, adolescents, and families.," *Pediatrics*, vol. 127, no. 4, pp. 800–4, Apr. 2011.

[7] [7] L. H. Shaw and L. M. Gant, "In Defense of the Internet: The Relationship between Internet Communication and Depression, Loneliness, Self-Esteem, and Perceived Social Support," *CyberPsychology Behav.*, vol. 5, no. 2, pp. 157–171, Apr. 2002.

[8] [8] M. L. Ybarra, "Linkages between Depressive Symptomatology and Internet Harassment among Young Regular Internet Users," *CyberPsychology Behav.*, vol. 7, no. 2, pp. 247–257, Apr. 2004.

[9] [9] S. Hinduja and J. W. Patchin, "Bullying, Cyberbullying, and Suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, Jul. 2010.

[10] [10] C. Van Hee *et al.*, "Automatic detection and prevention of cyberbullying," *Int. Conf. Hum. Soc. Anal.*, pp. 13–18, 2015.

[11] [11] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils," *J. Child Psychol. Psychiatry*, vol. 49, no. 4, pp. 376–385, Apr. 2008.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[12] [12] H. Vandebosch and K. Van Cleemput, “Defining Cyberbullying: A Qualitative Research into the Perceptions of Youngsters,” *CyberPsychology Behav.*, vol. 11, no. 4, pp. 499–503, Aug. 2008.

[13] [13] R. M. Kowalski, S. Limber, and P. W. Agatston, *Cyberbullying : bullying in the digital age*. Wiley-Blackwell, 2012.

[14] [14] D. Maher, “Cyberbullying: an ethnographic case study of one Australian upper primary school class,” *Youth Stud. Aust.*, vol. 27, no. 4, pp. 50–58, Dec. 2008.

[15] [15] 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. Human Behav.*, vol. 63, pp. 433–443, Oct. 2016.

[16] [16] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, “Towards understanding cyberbullying behavior in a semi-anonymous social network,” in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 244–252.

[17] [17] F. Baltar and I. Brunet, “Social research 2.0: virtual snowball sampling method using Facebook,” *Internet Res.*, vol. 22, no. 1, pp. 57–74, Jan. 2012.

[18] [18] A. Kontostathis and A. Kontostathis, “ChatCoder: Toward the Tracking and Categorization of Internet Predators,” *PROC. TEXT Min. Work. 2009 HELD CONJUNCTION WITH NINTH SIAM Int. Conf. DATA Min. (SDM 2009). SPARKS, NV. MAY 2009.*, 2009.

[19] [19] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, “Detection of Harassment on Web 2.0,” *PROCEEDINGS OF THE CONTENT ANALYSIS IN THE WEB 2.0 (CAW2.0) WORKSHOP AT WWW2009.* .

[20] [20] T. Hastie, J. Friedman, and R. Tibshirani, “Overview of Supervised Learning,” Springer, New York, NY, 2001, pp. 9–40.

[21] [21] H. B. Barlow, “Unsupervised Learning,” *Neural Comput.*, vol. 1, no. 3, pp. 295–311, Sep. 1989.

[22] [22] M. Andriansyah *et al.*, “Cyberbullying comment classification on Indonesian Selebgram using support vector machine method,” in *2017 Second International Conference on Informatics and Computing (ICIC)*, 2017, pp. 1–5.

[23] [23] S. C. Eshan and M. S. Hasan, “An application of machine learning to detect abusive Bengali text,” in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1–6.

[24] [24] Noviantho, S. M. Isa, and L. Ashianti, “Cyberbullying classification using text mining,” in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, 2017, pp. 241–246.

[25] [25] H. Nurrahmi and D. Nurjanah, “Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility,” *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018–Janua, pp. 543–548, 2018.

[26] [26] 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 - SAM ’14*, 2014, pp. 3–6.

[27] [27] S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu, “Detection of cyberbullying on social media messages in Turkish,” in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 366–370.

[28] [28] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth, “A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research,” in *Proceedings of the 10th ACM Conference on Web Science - WebSci ’18*, 2018, pp. 33–36.

[29] [29] B. Haidar, M. Chamoun, and F. Yamout, “Cyberbullying Detection: A Survey on Multilingual Techniques,” in *2016 European Modelling Symposium (EMS)*, 2016, pp. 165–171.

[30] [30] F. Del Vigna, A. Cimino, and F. D. Orletta, “Hate Me, Hate Me Not: Hate Speech Detection on Facebook,” in *First Italian Conference on Cybersecurity (ITASEC17)*, 2017, no. January, pp. 86–95.

[31] [31] 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 - ICDCN ’16*, 2016, pp. 1–6.

[32] [32] 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)*, 2015, pp. 611–616.

[33] [33] Parsania, Vaishali, Navneet Bhalodiya, and N. N. Jani. "Applying naïve bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis." (2014).

[34] [34] K. D. Gorro, M. J. G. Sabellano, K. Gorro, C. Maderazo, and K. Capao, “Classification of Cyberbullying in Facebook Using Selenium and SVM,” in *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, 2018, pp. 183–186.

[35] [35] Dehue, F., Bolman, C. & Völlink, T. 2008. Cyberbullying: Youngsters' experiences and parental perception. CyberPsychology & Behavior, 11, 217-223.

[36] [36] R., Broeren, S., Van De Looij–Jansen, P. M., De Waart, F. G. & Raat, H. 2014. Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents. PloS one, 9, e94026.

[37] [37] Shariff, S. & Patchin, J. W. 2009. Confronting cyber-bullying, Cambridge University Press.

[38] [38] Shariff, S. 2008. Cyber-bullying: Issues and solutions for the school, the classroom and the home, Routledge.

[39] [39] Campbell, M. A. 2005. Cyber bullying: An old problem in a new guise? Australian Journal of Guidance and Counselling, 15, 68-76.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500