Detection of Cyberbullying on Social Media Using Machine Learning

Vidhya.S[1], Dheivani.A[2], Gopinath.S[3],Manoj Kumar.S[4] ,Sobiya.S [5].
Assistant Professor[1],Department of Computer Science.
UG Scholar[2,3,4,5], Department of Computer Science.
Angel college of engineering and technology, Tirupur -641665.

Abstract—cyberbullying was confined to boundaries and has now moved online. Harassment is one form of cyberbullying. Social media cyberbullying incidents are escalating. Insulting words are dynamic, and the same word can have several meanings according to the context. Simply because a comment contains such a word, it cannot be classified as bullying. Hence, Labeling comments, and keyword spotting techniques are inadequate. Other languages have addressed this issue using lexical databases such as WordNet which provides synonyms and homonyms of words. Since there is no proper lexical database developed for the English language, detecting a word as bullying is a challenge. Therefore, we used rules to overcome this issue. Twitter comments with profane words were collected, outliers were removed, and the remaining tweets were pre-processed. These rules were used for feature extraction. Afterward, we applied Support Vector Machine (SVM), K-nearest neighbor (KNN), and Naïve Bayes algorithms. The results show that SVM with an RBF kernel performs better with an F1-score of 91%. The novelty of this research is the focus on English language cyberbully detection which has not been addressed before.

Keywords—KNN, SVM, RBF, Cyberbully, Social Media, Machine learning, Personal attack, Hate speech, Feature extraction

I. INTRODUCTION

Cyberbullying or cyberharassment is a form of bullying or harassment using electronic means.
Cyberbullying is also known as online bullying. It has become increasingly common, especially among teenagers, as digital technology has advanced. Cyberbullying or harassing others on the internet and other digital spaces, particularly on social media sites. Harmful bullying behavior can include posting rumors, threats, sexual remarks, a victim's personal information, or pejorative labels (i.e. hate speech). Bullying or harassment can be identified by repeated behavior and an intent to harm. Victims of cyberbullying may experience lower self-esteem, increased suicidal ideation, and various negative emotional responses, including being scared, frustrated, angry, or depressed.

Awareness in the United States has risen in the 2010s, due in part to high-profile cases. Several US states and other countries have passed laws to combat cyberbullying. Some are designed to specifically target teen cyberbullying, while others extend from the scope of physical harassment. Cyber harassment, reports are usually filed beginning with police. The laws differ by area or state.

Research has demonstrated several many consequences of cyberbullying victimization. Specific statistics on the negative effects of cyberbullying differ by country and other demographics. Some researchers point out there could be some way to use modern computer techniques to determine and stop cyberbullying. Internet trolling is a common form of bullying that takes place in an online community (such as online gaming or social media) to elicit a reaction or disruption, or simply just for someone's amusement. Cybers talking is another form of bullying or harassment that uses electronic communications to stalk a victim; this may pose a credible threat to the victim. Not all negative comments on social media can be attributed to cyberbullying. Research suggests that there are also interactions online that result in peer pressure, which can have a negative, positive, or neutral impact on those involved.

a) Sentiment analysis

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to the voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence, or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as enjoyment, anger, disgust, sadness, fear, and surprise. Precursors to sentimental analysis include the General Inquirer, which provided hints toward quantifying patterns in text and, separately, psychological research that examined a person's psychological state based on analysis of their verbal behavior. Subsequently, the method described in a patent by Volcani and Fogel looked specifically at sentiment and identified individual words and phrases in the text concerning for to different emotional scales. A current system based on their work, called Effect Check, presents synonyms that can be used to increase or decrease

the level of evoked emotion in each scale.

Many other subsequent efforts were less sophisticated, using a mere polar view of sentiment, from positive to negative, such as work by Turney, and Pangwho applied different methods for detecting the polarity of product reviews and movie reviews respectively. This work is at the document level. One can also classify a document's polarity on a multi-way scale, which was attempted by Pang and Snyder among others: Pang and Lee expanded the basic task of classifying a movie review as either positive or negative to predict star ratings on either 3- or a 4-star scale, while Snyder performed an in-depth analysis of restaurant and hotel reviews, predicting ratings for various things and aspect of the given restaurant, such as the food and atmosphere of environments (on a five-star scale).

The First steps to bringing together types in specific of approaches to learning, lexical, knowledge-based, etc.—were taken in the 2004 AAAI Spring Linguists Symposium, computer scientists, and other interested researchers some of the research linked to data systematic computational research in subjectivity, and sentiment in text. In one of the statistical classification methods, the neutral text is assumed and ignored in neutral classes that lie near the boundary of the binary classifier, several researchers suggesting in every polarity problem, three categories must be identified. Moreover, MAX entropy and SVM is the specific classifier and it proving can benefit from the introduction of a neutral class and improve the overall accuracy of this classification. There are in principle two ways for operating with a neutral class. the algorithm is to be proceeded by first identifying the neutral language and filtering it out and then assessing the rest in terms of positive and negative sentiments, and it can build a three-way classification in one step. This second approach often involves overall categories (e.g. naive Bayes classifiers as implemented by the NLTK). Whether and how to use a neutral class depends on the nature of the data: if the data is clustered into neutral, negative, and positive language, it makes sense neutral language can be filtered out and focus the polarity between positive and negative sentiments. If, in contrast, the data are mostly neutral with small deviations towards the positive and negative affect, this strategy would make it very harder to clearly distinguish between the two poles.

A different method for determining sentiment is the use of a scaling system and words commonly associated with having a negative, neutral, or positive sentiment with them are given an associated number up to −10 to +simply from 0 to a positive upper limit such as +4. This makes it possible to adjust the sentiment of a given term relative to its environment (using the level of the sentence). When a piece of unstructured text is analyzed using natural language processing(NLP), each concept in the specified environment is given a score based on the way sentiment words relate to the concept and its associated score of the relevant topic. This allows movement to in sophisticated understanding of sentiment because it is now possible to sentiment in adjust the value of a concept relative to modifications that may surround it. Words, for example, that intensify, relax, or negate the sentiment expressed by also the concept can affect its score. Alternatively, texts can be given a positive and negative sentiment strength score if the goal is to determine a text rather than the overall polarity and strength of the text. There are various other types of analysis - Aspect Based on sentiment analysis, Grading sentiment analysis (positive, negative, neutral), and Multilingual sentiment analysis in the detection of emotions.

II. LITERATURE SURVEY

A System with the detection of cyberbullying on social media involves brief classification and extraction approaches. But further developments in the system lead to the development of efficient processors and classifiers. So it is better to study the recent developments in technology to know more features and possibilities. A detailed study has been made to identify the various techniques used and to analyze their merits and demerits. Following are the a few techniques that have been surveyed for improving the authentication of the characters in the process.

2.2 A CYBERNETIC FRAMEWORK TO ARTICULATE THE ORGANIZATIONAL COMPLEXITY OF USERS' INTERACTIONS WITH THE JIGSAW TECHNIQUE IN AN OPEN SIM STANDALONE SERVER

The purpose of this study was the utilization of an Open Sim standalone server in order to articulate and propose a theoretical cybernetic framework that delineated from the requirements of: (a) the systemic organizational structure of a learning process through the Viable System Model (VSM) and (b) the exploratory collaborative knowledge construction according to the "Jigsaw" learning technique for handling the interior organizational complexity of cyber entities' (avatars) interactions in this virtual world. The community empowerment may initially be amplified from the implementation of this cybernetic framework for enhancing the dynamic and interactive dimensions of users' (students and instructors) presence in Open Sim according to the processes of cohesion, coordination, and organizational processing that are quite complex during their first introduction.The construction of a well-defined model for developing an organizational framework and its implementation in VWs was the main contribution of this research. The idea of using collaborative learning techniques, such as the "Jigsaw" with its intermediates in "open source" virtual environments basically on design principles and the creation of learning scenarios, it is today imperative. As for the contribution from the utilization of the Jigsaw it was finally identified that: (a) Supported and highlighted cyber entities' (i.e. students and teachers) learning process in order to expand the interactive action through their search for other information sources beyond to the real, escaping from the traditional weakness created by the lack of an organizational framework for teaching process to address on the configuration of more complex interactions. (b) Applied to an open-ended environment that allows designing different learning approaches with Open Sim which until now had not been treated as an educational tool in university-level courses. With the evolving field of technological innovation for

educators or students, it provides learning opportunities that contribute to the knowledge management. Meanwhile, it changes how people adapt to a complex technologically-advanced environment, learn new tasks, and acquire new e-skills and access to educational resources on a global scale, leading to new strategies for learning, beyond statutory and formal contexts.

The utilization of knowledge resources, if there are properly well-organized, can improve students' performances in virtual environments. In this case, of course, if it wants to decipher the factors of success in these projects that must meet certain conditions, such as: (i) The connection of a collective goal of working for the needs of real life 22 (ii) The ease utilization of learning tools for a proper technical and organizational infrastructure of the system (iii) The formation of multiple communication channels for knowledge transfer and students' motivation. The overall conclusion for (a) question showed that the adoption of innovative environments for "knowledge management" can produce basic design principles of a digital environment, which should be adapted collaborative learning principles, the facilitation of the knowledge's co-construction as a collective process, and even more of the collective sharing knowledge configuring among the members of each group. Regarding the role of virtual environments in this multi-level shape of collective interactions, it is truly understandable that the technical capabilities of the environment facilitated by the construction of shared meanings and artifacts which cyber entities have to use in these grids, it was confidently a simulation of the individual collective work. Furthermore, it is worth noting participation in a group does not mean that individual representatives identified easier with collaborative activities. As regards the (b) question, the original findings of a strategic framework for further development as a "participatory cybernetic framework" in conjunction with the "Jigsaw" that constructed as an organizational plan in conjunction with the Jigsaw technique that seemed at least initially can be responding effectively on:

- Making useful the management of the learning process and increasing the mobility of interaction between students.
- Improving the quality and effectiveness of education and training.
- Fostering innovation and creativity and entrepreneurship at all levels of education and training sessions.

To sum up, the reasons for (c) question showed that the dynamics of cooperation from such an action framework could bring up a new dimension among the facts of cyber entities with the "Jigsaw" technique and the "cybernetics mechanisms" of a "viable" learning process, improving the quality and effectiveness of the e-Education status quo and training process (among others) emplacing on: (a) Participating users in the development of the training workshop where they employed. (b) Generating new knowledge and its innovation through student's involvement in Open Sim. (c) Evacuating students "intrinsic cognitive overload" as students become autonomous leaders in their professional development throughout their careers. In this study we proposed briefly a "cybernetic" improvement research framework for the reeducation, but it occurred that needs to develop or provided more instructional affordances through for the process of lifelong learning and continuing education. Future works may illustrate on this educational cybernetic practice some evaluation fragments (with formative or summative evaluation form) to ensure the high quality of this "cybernetic" framework and to provide the initial groundwork for Higher Education courses in a more systemic way. These establishments currently can enforce the continuous professional development of teachers and trainers, making the teaching plan as an attractive choice transmitting students' transversal skills/abilities. It is an optimization limn problem, which falls within the area of operations management of financial resources for educational research. In particular, this problem concerns the production of goods and services and involves the responsibility of ensuring that educational activities are effective, both in terms of the minimum use of resources required and the complete satisfaction of students' needs.[1]

2.3 AUTOMATED DETECTION OF CYBERBULLYING USING MACHINE LEARNING

Niraj Nirmal, Pranil Sable Increasing the use of Internet and facilitating access to online communities such as social media have led to the emergence of cybercrime. Cyberbullying is very common now a days. which have no tracking like it may harm any individual, business, society, country in past few days it seems that riots were happened due to some statement used by one community on another its important to identify such content which spreads hate or harm community text processing, NLP (natural language processing) is an emerging field with the help of NLP and machine learning algorithms such as naive bayes, random forest, SVM we are going to identify cyberbullying in facebook. Objectives of this implementation written in objective section. Image character with the help of OCR will be done by us to find image - based cyberbullying the impact on individual basis thus will be checked on dummy system. Machine learning and natural language processing techniques to identify the characteristics of a cyberbullying exchange and automatically detect cyberbullying by matching textual data to the identified traits. On the basis of our extensive literature review, we categorise existing approaches into 4 main classes, namely supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. Supervised learning-based approaches typically use classifiers such as SVM and Naïve Bayes to develop predictive models for cyberbullying detection.

The validity and accuracy of the predictive models to detect cyberbullying on facebook in this case primarily based on the correct psychometric categorization of the text. In future it is intended to improve the system developed by use more accurate dataset and to detect the cyberbullying or not. We also apply other machine learning algorithm and check the accuracy of models. Higher accuracy model will help to detect more accurate bullying. Another interesting direction for future work would be the detection of fine-grained cyberbullying categories such as threats, curses and expressions of racism and hate. When applied in a cascaded model, the system could find severe

cases of cyberbullying with high precision. This would be particularly interesting for monitoring purposes. Additionally, our dataset allows for detection of participant roles typically involved in cyberbullying. The goal of this project is to the automatic detection of cyberbullying- related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary. However, these posts could just as well indicate that cyberbullying is going on. The main aim of this project is that it presents a system to automatically detect signals of cyberbullying on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders.[2]

2.4 WEB FILTERING AND CENSORING

Thomas M. Chen and Victoria Wang, International security researchers argued that the !lter was inaccurate and contained vulnerabilities potentially exposing PCs to security threats. Free speech advocates expressed concern that the government could use the !lter to monitor users' online activities and block politically sensitive websites. And the US government urged the Chinese Ministry of Industry and Information Technology and Ministry of Commerce to revoke the Green Dam requirement on the basis of free trade. In response to the controversy, the Chinese government inventively "delayed" the requirement except for PCs used in schools, cyber cafes, and other public access locations. Around the same time, following the controversial reelection of President Mahmoud Ahmadinejad in Iran, critics accused the regime of blocking certain websites such as Facebook and YouTube—which had been used to post confrontations with the police—as well as sites af!liated with the opposition leader. The Iranian government was also suspected of monitoring Internet usage to track down election protesters. In January 2010, public attention was again drawn to the issue of Web censoring when Google.Cn decided to stop complying with Chinese government requirements to censor search results related to politically and socially sensitive issues. Google reached its decision, considered long overdue by some, in response to attacks by Chinese hackers on Gmail accounts of Chinese human rights activists during the previous month. US Secretary of State Hillary Clinton publicly praised Google and called for a global end to Web censoring, prompting a critical response from the Chinese government.

Given the Chinese government's effective covert Web censoring program, its mandate for the desktop Green Dam filter was puzzling. Although the !lter would have been installed on every PC, it could easily be disabled. Sufficiently motivated individuals can circumvent a Web !lteror example, they can bypass a URL blacklist by going directly to the server's IP address. In the case of China, Golden Shield does not attempt to read or censor any encrypted traffic such as used in virtual private networks (VPNs). Virtually all foreign businesses in China depend on VPNs, and blocking them would unacceptably impact commerce. Chinese citizens also use various proxy tools, including Garden Networks' Gunnel These products use a combination of proxies, encryption, and onion routing to offer anonymized and uncensored Web access. If Web! altering is easy to circumvent, then what purpose does it serve? In the case of China, the apparent goal is to impose Signiant obstacles on the country's 350 million Internet users to discourage the majority from accessing certain foreign websites and thereby keep public attention on tightly regulated domestic sites. Only a small minority have the technical know-how to !Nd ways around government censorship.[3]

III. EXISTING SYSTEM

Machine learning (ML) is a data analysis that automates analytical model building. ML algorithms are categorized as supervised or unsupervised. Supervised ML algorithms have been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. Unsupervised ML algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies and describe a hidden structure from unlabeled data. The problem with unsupervised ML is that they may overlap and learn to localize texts with minimal unsupervised algorithms. Many researchers have used supervised learning approaches on data related to publicly released corpora. **NB** classifiers as supervised learning models are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence between the features. They are the simplest Bayesian network models. NB often relies on the bag of words presentation of a document, where it collects the most used words neglecting other infrequent words. The bag of words depends on the feature extraction method to provide the classification of some data. Furthermore, NB has a language modeling text as a representation of unigram, bigram, and n-gram and tests the probability of the query with a specific document.

IV.PROPOSED METHODOLOGY

Support-Vector Machines (SVMs) classifier are used as supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, a SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). The most important models for SVM text classifications are Linear and Radial Basis functions. Linear classification tends to train the data. or identifying insult were defined, collected comments were preprocessed, features were extracted, and several machine learning algorithms were used to train the model. Finally, results were compared and evaluated to identify the best suited method.

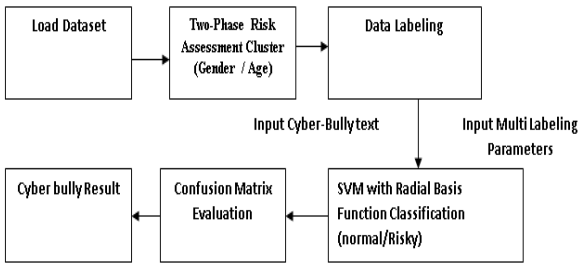

Fig 1. Methodology

IV.MODULES

1.DATA ANALYSIS

This module is used to load XML based data set for cluster to find the most likely cluster for each given data item. More precisely, to determine K clusters over a set of data items, we have to define K probability distributions, each one representing the likelihoods of data items to belong to a given cluster. In our setting, in the first phase, membership probabilities are computed based on the values of GI features. Then, based on these likelihoods, each user is associated with the group that better fits his/her GI features that is, the one with the highest membership probability. In the second phase, users of the same group are further clustered according to their behavioral features.

2.GROUP IDENTIFICATION

This module is aim of the first clustering is to group users for which similar behaviour are expected like male /female gender based clustering the best cluster is computed in a deterministic way ,this is, each item is assigned to the unique cluster. In contrast, soft clustering computes, for each item and every available cluster, the membership probability. This modules helps target user we are more interested in having his/her cluster membership probability.

3.SVM RADIAL BASIS FUNCTION CLASSIFICATION

We recall that our risk assessment is composed of two phases first aim at organizing users according to group identification features and then according to behavioural features .Regardless of the features ,regardless of the features taken into account, in both these phases we make use of the same clustering algorithm.a user behavioural profile able to catch those user's activities and inter action that are considered meaningful for thr risk assessment using our proposed model. The second issue regards how to model a 'normal behaviour'/risky (cyberbully-user)using different hyber-plane labels based on RBF kernel values.

4.USER RISK SCORE

User that diverge from normal behaviour. These deviations are actually captured by the membership probabilities compute in the second clustering phase.More precisely, a high membership probability value implies that the target user fit well one of the behaviour emerged from the group he/she belongs to. The risk score associated with a target user u is defined as the inverse of the highest among membership probability values resulting by the second clustering phase. The extracted features will be used to analyze for classification as similar to normal users or risky cyber bully user.

V.CONCLUTION

Cyberbullying social media comments in the English language using a hybrid method, which is a combination of a rule - based approach and machine learning. The usage of social media is dramatically increasing, and certain people have turned them into a platform to bully other people. There had been reported incidents where people were humiliated for their appearance using insulting comments on various social media platforms. Up to now, there is no proper way of removing these insulting statements other than reporting them. Even after reporting, these insulting statements might not be removed because of the absence of sufficient English language translators. To address the problem of the lack of human resources in the form of language interpreters, we built a text analytics model using rules and machine learning algorithms. In order to identify harassment in a social media text, we used five rules. Rules that were appropriate for the English language were produced with an expert decision. These rules were different from that of research on the Indonesian language. Then the features were fed into several machine-learning algorithms to learn patterns. The results were compared to identify the best method. The testing results had an F1-score of 91% for SVM with RBF kernel.Therefore, we believe that this research will be useful to automatically detect English language cyberbullying text. When the data set was above 600 records, the accuracy was 93% and F1-score was 91%. When data set was below 200 records the accuracy was 85%. This shows that when the number of records increase, the accuracy and F1-score increases. Therefore, we can conclude that this method is much suitable for a large dataset than a small dataset. More frequently people use English characters to 56 write English words on social media. In our research, we have only considered socialmedia comments that are written using pure English language. Therefore, as future work, we will be expanding the data corpus to include English language comments that are written using English characters as well. Further, the model could be improved to obtain a high precision with increasing recall value with a larger data corpus.Further more, the English bad word list could be expanded with the help from participants in the labelling system and with the expert knowledge. Moreover, the manual labelling process could be automated to make the process more efficient.

VI.REFERENCE

[1]. Tavani, Herman. T., "Introduction to Cybernetics: Concepts, Perspectives,and Methodological Frameworks", In H. T. Tavani, ethics and Technology:Controversaries, questions, and Strategies for ethical Computing, river University – Fourth Edition, Wiley, pp 1-2, 2013.

[2]. S. Salawau, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A survey," Vol. 3045, no c, pp 1-20, 2017.

[3]. Internet Monitoring and Web Filtering Solutions","PEARLSOFTWARE,2015.Online.Avaliable:http ://www.pearlsoftware.com/solutions/cyber- bullying in schools.html. [Accessed Feb 20, 2020]

[4]. K. Reynolds, "Using Machine Learning to Detect Cyberbullying", 2012.

[5]. V. Nahar, X. Li and C. Pang, "An effective Approach for Cyberbullying Detection," in Communication in Information Science and Management Engineering, May 2013.

[6]. Chen, Y., Zhou, Y., Zhu, s. and Xu, H., "Detecting Offensive Language in Social Media to Protect Adolescent Online Saftey", In privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing(SocialCom), pp 71-80, 2012.

[7]. B.Sri Nandhinia, and J.I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Appearance Patterns", in Knowledge and Smart Technology(KST), 2017 9th International Conference on, pages 242-247, IEEE, 2017
.

[8]. Walisa Romsaiyud, Kodchakornna Nakornphanom, Pimpaka Prasertslip, Piyapon Nurarak, and Pirom konglerd," Automated Cyberbullying Detection Using Clustering Appearance Patterns", in Knowledge and Smart Technology(KST), 2017 9th International Conference on, pages 242-247, IEEE, 2017.

[9]. Dipika Jiandani, Riddhi Karkera, Megha Manglani, Mohit Ahuja, Mrs.Abha Tewari, "Comparative Analysis of Different Machine Learning Algorithms to Detect Cyber-bullying on Facebook", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume6Issue IV, April 2018, pp. 2322- 2328.

[10]. Cristina Bosco and Viviana Patti and Andrea Bolioli, "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti–TUT, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI2015), pp. 4158-4162.