Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 181 (2021) 605–611

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2020

Accurate Cyberbullying Detection and Prevention on Social Media

Andrea Perera[a*], Pumudu Fernando[b]

[a]*AndreaPera,Department of Computer Science,Informatics Institute of Technology,Colombo 06, Sri Lanka*
[b]*Pumudu Fernando,Department of Computer Science,Informatics Institute of Technology,Colombo 06, Sri Lanka*

Abstract

The usage of digital/social media is increasing day by day with the advancement of technology. People in the twenty-first century are being raised in an internet-enabled world with social media. Communication has been just one button click. Even though there are plenty of opportunities with digital media people tend to misuse it. People spread hatred toward a person in social networking. Cyberbullying affects people in different aspects. It doesn't affect only for health, there are more different aspects which will lead life to a threat. Cyberbullying is a worldwide modern phenomenon which humans cannot avoid hundred percent but can be prevented. Most existing solutions have shown techniques/approaches to detect cyberbullying, but they are not freely available for end-users to use. They haven't considered the evolution of language which makes a big impact on cyberbullying text.
This paper describes a system for automatic detection and prevention cyberbullying considering the main characteristics of cyberbullying such as Intention to harm an individual, Repeatedly and over time and using abusive curl language or hate speech using supervised machine learning. The system relies on the detection of cyberbullying text along with the themes/categories associated with cyberbullying such as racist, sexual, physical mean, swear and other, using support vector machines and Logistic regression. The author of this research presents a new hypothesis for cyberbullying detection that the circumstances and usage of texting and its language have changed by time. Most of the studies have considered calling someone stupid, ugly and idiot as cyberbullying. Things have changed, such words may or may not always be a bullying incident. If a person wants intentionally to harm an individual, they will use extreme words. In addition to traditional feature extraction techniques like Term Frequency–Inverse Document Frequency (TF-IDF), N-gram and profanity along with sentiment analysis increases the accuracy of the system. Evaluated the proposed system using Recall, Precision and F1-score.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: andrea.2016432@iit.ac.lk

1877-0509
10.1016/j.procs.2021.01.207

Keywords: Cyberbullying; Supervised machine learning; Natural language processing; Social media; Health problems.

1. Introduction

Cyberbullying was not in our vocabulary a decade ago. The invention of the World Wide Web(WWW) made a huge impact on social media as they were fast, they didn't need to be physically present only needed a digital device [1]. Cyberbullying is aggressive, abusive behavior towards a person or a group of people intentionally, repeatedly harming them by spending offensive content or engaging in other forms of social violence through using digital media [2],[3],[4]. Cyberbullying is a huge phenomenon among teenagers as a victim or predator or bystander [5]. Hinduja and Patchin show that 36.5% of students have experienced cyberbullying at least once in their lifetime. This shows that mean or hurtful comments online had the highest number out of all the other types [6]. A survey carried out in the USA 1,501 youths aged 10–17 years, it has been found that around 12% have admitted that they were being abusive to someone online and 4% were victims of aggressiveness, while around 3% were both victims and have bullied someone [7].

In Sri Lanka survey carried out by Cyber Crimes Division (CID) which is the Law enforcing Authority shows that more than 1000 cyberbullying cases were reported. Nearly all survey participants said they knew people harassed online and around 90% of Students at the university have been exposed to cyberbullying. Facebook was the platform where Sri Lankans have seen 80% of cyberbullying.65% of university students have posted embarrassing videos or photos.15% was posting private information online, 9% spreading false information of people and lies and 2% posting abusive content [8].

The biggest challenge is that cyberbullying can quickly spread as there's a large audience and stay visible for a longer time. Cyberbullying is a modern phenomenon as victims have to face lots of consequences. It has caused major issues to mental and physical health. Schneider et al show the relationship between victimization and 5 types of psychological distress from the MetroWest Adolescent Health Survey. data collected from more than 20000 students. The depressive symptoms (34%) and self-injuries (24%) were highest among the victims of Cyberbullying. Most of the cases won't be reposted to a responsible party to act against it [9]. Facebook, Instagram and Twitter also have addressed this problem. A victim can report and block the predator. Facebook Bullying prevention Hub is available now for Teens, Parents and Educators to support and help [10]. Even Instagram has introduced a new feature powered by AI "Twitter's online abuse hub giving a chance to the users to think before they Tweet. There are lots of tools and software to block the predator if there are abuse messages. Most of the social media today prohibit the use of abusive and insulting comments. These ways are not efficient as it's time consuming to monitor or to inspect all the reported posts daily. Many researchers in our field introduced many approaches to detect cyberbullying in social media which will be discussed in the next section.

2. Related Work

Yin et al.[11], used a supervised learning approach to detect harassment from three different social websites. Used Kongregate (chat-style community), Slashdot and Myspace (discussion style communities) as a dataset along with the content features, sentiment features and contextual features of documents. They used a libSVM with the linear kernel as a classification tool. The performance of TFIDF weighting was better than n-gram and Foul Language.

Based on three different features, a profanity dictionary, Levenshtein Edit Distance and Bag-of-words [12] established three profanity detection systems. The profanity dictionary was based on a user-compiled list on phorum.com and noswearing.com. In addition to an edit distance calculator, the second system used a list of swear words to adjust misspellings. The system checks the words against an English dictionary and a list of names to reduce false positives. As an example, an edit distance calculator will match' shirt' with the profane term ' shit' and label ' shirt' as an offensive word, but the program will classify the word 'shirt' as not profanity by checking the dictionary.

The third detection system used SVM (Support Vector Machine) classifiers along with word stems and bigrams as features. By running a series of tests using the three detection systems in various combinations, they produced the best overall result using a configuration combining the results of all three systems in an "OR" operation. Example: if a comment was identified as profanity by any of the three systems and the most accurate combination used the SVM-based system "AND" either the profanity list or the Levenshtein distance-based system.

Squicciarini et al. [13] used personal, social networking and content-specific features with a Decision Tree classifier to identify bullies on Myspace and spring.me and built a set of rules to assess if the cyberbullying activity of a user is initiated by the actions of another bully.

Chavan and Shylaja [14] also produced a score signifying other users the probability that a statement could be offensive. They achieved a 4 percent increase in accuracy using a dataset from Kaggle10 by using a variety of features including skip-grams and integrating the outcomes of Support Vector Machine and Logistic Regression classifiers.

Most of the existing systems have focused on ways or techniques detecting cyberbullying but don't have a system which end-users can directly control or use and haven't focused on sarcastic content. The proposed solution is automatically detecting cyberbullying incidents and prevention, does not consider sarcastic content as cyberbullying and identifying the themes or categories associated with cyberbullying.

Most of the studies have considered calling someone stupid, ugly and idiot as cyberbullying. The author believes that things have changed, presented a new hypothesis such words (stupid, ugly and idiot) may or may not always be a bullying incident. if a person wants intentionally to harm an individual, they will use extreme words such as Cunt, Bitch, Dickhead, etc. not just idiot, stupid. This proposed solution will help cyberbullying victims to take decision or action against the bullying person as it will be automatically labeled as cyberbullying or not.

3. Background

In most of the researchers and solutions, they have only considered one main characteristic of cyberbullying which is detecting abusive language/insult words or identifying the role of cyberbullying. Cyberbullying cannot be detected by considering only one characteristic. Cyberbullying detection is difficult and different than just detecting aggressive content. Detecting abusive language/insult words or hate speech is not a valid reason to show that it is cyberbullying. Some edge cases are shown below.

- "**This bog is stupid I don't like it**" which only includes an insult word or profanity which not cyberbullying.
- "**You are a cunt**" is a case which is a cyberbullying and profanity/insult word along with a second person(you) or third person (She, He, They, It) or a person's name.
- "**You are not an idiot**", profanity/insult word along with a second person(you) or third person (She, He, They, It) or a person's name but it's not a cyberbullying incident.
- "**I hate your fucking phone sometimes but it so useful**" which used a profanity word to express more or exaggerate or act as an intensifier (So, very).

Most of the studies and researches used a single message to indicate cyberbullying and the current dataset which is used to train the model haven't correctly labeled the sarcastic text. The author proposed that additional aspects need to be considered when it comes to detecting cyberbullying. There are other characteristics to consider whether this is cyberbullying or not. They are intending to harm an individual, repeatedly and over time, an imbalance of power in a situation where the victim cannot defend themselves.

- "**Why am I so stupid**" this example event though there is an insulting word this doesn't harm an individual, the person is expressing his feeling using is an insulting word which can be considered as non-cyberbullying.

As an example, people do use bad words when texting and the other person doesn't feel bad about it or consider it as a joke.

- **Person 01: You did fucking well on the exam!**
  Person 02: Thanks, m8(mate)

In order to create accurate cyberbullying detection, we have to develop a system that will adapt to the language changes. The structure of the sentence changed with time. As an example, "How do you do?" has become "How's it going?", using "u" for "You" and using abbreviations like STFU, ILY, etc. The author of this research makes up with

a new hypothesis for cyberbullying detection that in cyberbullying cases people tend to use extreme words rather than idiot, stupid, etc.

The above few examples illustrate the difficulty of the cyberbullying identification task using standard natural language processing techniques, which involves an analysis of all the phrase textual information.

4. Proposed Solution

This section shows the methodology of the proposed solution. It is shown in the Figure 1.

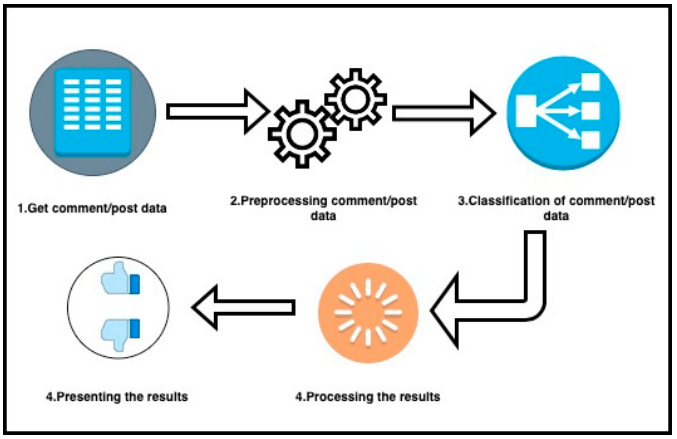

Fig. 1. proposed system

The author of this research used a Twitter dataset, where globally people have seen the most amount of cyberbullying, retrieve it from Internet Archive [15], which is an American open source digital library of websites, software applications/games, music, movies/videos, and millions of books. The author only stored the relevant data from the Twitter JSON object such as the text of the tweet, tweet ID, in reply to status ID and retweeted status ID, used a SQL database to store them.

The noise of the text and unnecessary features can negatively affect the overall performance of the model. Therefore. For each sentence web links, emojis, extra punctuation and unwanted, repeated characters were removed. Next, converted contractions (“I’m”, “there’re”) and abbreviations into converting them to formal language. The author adds the most common abbreviation to a list and tokenizes each sentence, to extract individual words in a sentence. Search abbreviation words in the list and replace them with the formal language word. Removed some of the pronouns from the English stop words list such as “you”,” she”,” they” etc. Converted all the text into lowercase in order to reduce the document space problem.

4.1. Features

The author selected some of the best Content-based Features and Sentiment-based Features as machine learning algorithms can’t handle all the features. Content-based Features are based on the contents of the comments. Sentiment-based Features are based on the emotion of the text.

- TF-IDF (Term Frequency, Inverse Document Frequency)- By using TFIDF which can measure the importance of words in a document and Common words such as “is”, “am” do not affect the results due to IDF.
- Profanity along with Pronoun- Most of the cyberbullying content found profanity. Profanity alone itself is not a key aspect to detect cyberbullying. In some studies, hypothesizes that cyberbullying text contains a swear word/insult word along with the second person(you) or third person (She, He, They, It) or a person’s name [16],[17]. The author checked whether the text has a profanity word alone with a pronoun.

Table 1. Labeling the text

Text	"She"	"her"/name	"She is"	Profanity word	Label
1	Present	Present	Not present	Present	Cyberbullying
2	Present	Present	Not present	Not present	Non-Cyberbullying
3	Not present	Not present	Present	Present	Cyberbullying

- Frequency of cyberbullying themes/categories associated with cyberbullying such as racist, sexual, physical mean, swear and other, the author has manually compiled a list of words associated with racist, sexual, physical mean and swear. Author of this research will be taking a percentage of the associated words based on the word count of the full sentence.as an example " you are a black nigger", word count associated with race is 2 and full word count in the sentence is 3, the percentage is 67% ((2/3) *100)

- Sentiment analysis-is analyzing the whole sentence and understanding the perspective it expresses.it can be positive or negative meaning which is called polarity. the polarity score signs as a positive, neutral, or negative. AFFINN lexicon was used for this purpose. It comprises over 3,300 words, associated with a polarity score for each word. 51.5% of the labeled cyberbullying posts were negative.

4.2. Classifier

Author used Support Vector Machines (SVM), A well-known efficient binary classifier to train our model. Logistic regression was used to select the best combination of features. SVM algorithm, training data is used to learn a classification function. It can classify new data not previously seen in one of the two categories. It separates the training data set into two categories using a large hyperplane. Logistic regression is a linear classifier that predicts the probabilities, not classes [18].

4.3. Data labelling

Training the model is an essential phase of the supervised classification task of the SVM. To train the model, it needs manually labeled texts whether they are cyberbullying or not. The author manually labeled 1000 texts and are still working on labeling. The manually labeled text consists of sarcastic text which is not cyberbullying.

4.4. User Interface

The Interface is built with user-friendliness and simplicity. Users can add a text post or comment. To build a web application will be using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and Bootstrap.

5. Experimental Results

30% of the manually labeled Twitter dataset which was randomly selected was used for testing purposes, while the remaining 70% was applied to train the model. The evaluation parameters were accuracy, precision, recall and F1-score.

Table 2. Result on test dataset using various features with SVM classifier

Feature	Accuracy	Precision	Recall	F1-score
TF-IDF	74.50%	74%	74%	74%
Sentiment analysis	69%	69%	68%	68%
Profanity	56%	32%	56%	41%
TF-IDF +Sentiment analysis	75.17%	75%	75%	75%

- The use of TF-IDF +Sentiment analysis provides as expected the best result in terms of accuracy, precision, recall, and, thereby, F1-measure.
- TF-IDF still performs individually much better than other individual features.
- TF-IDF performs much better when the dataset is large compared to TF-IDF +Sentiment analysis feature.

6. Conclusion and future work

In this paper, the author has presented the proposed solution which uses natural language processing techniques and supervised machine learning to detect cyberbullying accurately along with the themes/categories associated with cyberbullying such as racist, sexual, physical mean, swear and other. In order to create an accurate cyberbullying detection, the author has proposed a system that will adapt to the language changes with a new hypothesis. The proposed solution does not detect sarcastic text as cyberbullying. The proposed solution resulted in 74.50% accuracy along with 74% precision, 74% recall and 74% F1 Score. As this research is still ongoing, the author is working on getting higher accuracy. In future studies, other user features and network features also can be accountable.

References

[1] C. E. Notar, S. Padgett, and J. Roden, "Cyberbullying: A Review of the Literature," *Univers. J. Educ. Res.*, p. 9, 2013.
[2] H. Rosa *et al.*, "Automatic cyberbullying detection: A systematic review," *Comput. Hum. Behav.*, vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/j.chb.2018.12.021.
[3] B. S. Nandhini and J. I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," *Procedia Comput. Sci.*, vol. 45, pp. 485–492, 2015, doi: 10.1016/j.procs.2015.03.085.
[4] S. A. Hemphill, A. Kotevski, and J. A. Heerde, "Longitudinal associations between cyber-bullying perpetration and victimization and problem behavior and mental health problems in young Australians," *Int. J. Public Health*, vol. 60, no. 2, pp. 227–237, Feb. 2015, doi: 10.1007/s00038-014-0644-9.
[5] A. Ioannou *et al.*, "From risk factors to detection and intervention: A metareview and practical proposal for research on cyberbullying," in *2017 IST-Africa Week Conference (IST-Africa)*, Windhoek, May 2017, pp. 1–8, doi: 10.23919/ISTAFRICA.2017.8102355.
[6] J. W. Patchin, "2019 Cyberbullying Data," *Cyberbullying Research Center*, Jul. 09, 2019. https://cyberbullying.org/2019-cyberbullying-data (accessed Oct. 11, 2019).
[7] A. A. Mazari, "Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies," in *2013 5th International Conference on Computer Science and Information Technology*, Amman, Jordan, Mar. 2013, pp. 126–133, doi: 10.1109/CSIT.2013.6588770.
[8] "Harassment Beyond Borders; Can Victims Be Protected By Cyber Bullying In Sri Lanka?," *Colombo Telegraph*, Apr. 15, 2019. https://www.colombotelegraph.com/index.php/harassment-beyond-borders-can-victims-be-protected-by-cyber-bullying-in-sri-lanka/ (accessed Apr. 17, 2020).
[9] S. K. Schneider, L. O'Donnell, A. Stueve, and R. W. S. Coulter, "Cyberbullying, School Bullying, and Psychological Distress: A Regional Census of High School Students," *Am. J. Public Health*, vol. 102, no. 1, pp. 171–177, Jan. 2012, doi: 10.2105/AJPH.2011.300308.
[10] "Safety Center," *Safety Center*. https://www.facebook.com/safetyv2 (accessed Oct. 14, 2019).
[11] D. Yin, Z. Xue, and L. Hong, "Detection of Harassment on Web 2.0," p. 8, 2019.
[12] S. O. Sood, E. F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 2, pp. 270–285, Feb. 2012, doi: 10.1002/asi.21690.
[13] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, Paris, France, 2015, pp. 280–285, doi: 10.1145/2808797.2809398.
[14] V. S. Chavan and Shylaja S S, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi, India, Aug. 2015, pp. 2354–2358, doi: 10.1109/ICACCI.2015.7275970.
[15] "Internet Archive Search: collection:twitterstream." https://archive.org/search.php?query=collection%3Atwitterstream&sort=-

publicdate&page=2 (accessed Jul. 23, 2020).
[16] F. Mishna, M. Saini, and S. Solomon, "Ongoing and online: Children and youth's perceptions of cyber bullying," *Child. Youth Serv. Rev.*, vol. 31, no. 12, pp. 1222–1228, Dec. 2009, doi: 10.1016/j.childyouth.2009.05.004.
[17] J. W. Patchin and S. Hinduja, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," *Youth Violence Juv. Justice*, vol. 4, no. 2, pp. 148–169, Apr. 2006, doi: 10.1177/1541204006286288.
[18] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.
[19] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm," *Medium*, May 28, 2020. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (accessed Jul. 26, 2020).