Empirical Analysis of Supervised Machine Learning Techniques for Cyberbullying Detection

Akshi Kumar, Shashwat Nayak and Navya Chandra

Abstract Cyberbullying is utilization of digital technology for targeting a person or a group in order to bully them socially and psychologically. Real-time social media platforms such as Instagram, Twitter, and YouTube have a large viewership, which serves as a fertile medium for such bullying activities. Instances of such harassment or intimidation are maximally found in the comments of a dynamic and an expressive medium like YouTube. This necessitates an adequate requisite to take relevant steps to find solutions for the detection and prevention of cyberbullying. The work presented in this paper focuses on the implementation of four supervised machine learning methodologies, namely Random Forest, k-Nearest Neighbor, Sequential Machine Optimization, and Naive Bayes in order to identify and detect the presence or absence of cyberbullying in YouTube video comments. The experimentation was carried out expending the Weka toolkit and utilizing the data gathered from comments obtained from YouTube videos involving core sensitive topics like race, culture, gender, sexuality, and physical attributes. The results are analyzed based on the measures like precision, accuracy, recall, and F-score and amongst the four techniques implemented, k-Nearest Neighbor is able to recognize the true positives with highest accuracy of around 83%. We also discuss various future research prospects for detection of cyberbullying.

Keywords Cyberbullying · Social media · Supervised machine learning
YouTube

A. Kumar (✉) · S. Nayak · N. Chandra
Delhi Technological University, New Delhi, Delhi, India
e-mail: akshikumar@dce.ac.in

S. Nayak
e-mail: shashwatnayak@outlook.com

N. Chandra
e-mail: navya.chandra2010@gmail.com

S. Bhattacharyya et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 56,
https://doi.org/10.1007/978-981-13-2354-6_24

1 Introduction

With the advent of Web 2.0 technologies, the use of the internet has increased tremendously [1] and its pervasive reach has some unintended consequences. One of the negative outcomes is cyberbullying, which is a major concern for society. Bullying taking place over devices like cell phones, tablets, and computers are known as cyberbullying. It can happen through messages and apps, or online through social networks, forums, or gaming portals where anyone can look at, participate in, or share information and content [2]. It comprises of but is not limited to posting, sending, or sharing undesirable, damaging, dishonest, or degrading content about someone else. It might include sharing personal details of a person causing embarrassment or humiliation. Some cases can lead to unlawful or criminal behavior. Harassment, denigration, flaming, impersonation, masquerading, pseudonyms, trolling, and cyberstalking are some of the common ways in which cyberbullying takes place [3].

Cyberbullying has an adverse effect both on the minds of the victim and the bully. A bully is someone who uses strength, force, or threat to intimidate or abuse another. This happens or is capable of happening over an extended period of time. A cyberbully, quite unlike a traditional bully can make use of the varied resources available on the internet to render himself anonymous. This boosts his confidence and many a times, he is also not able to gage the reaction of his victims who might similarly be unknown to him. This is also known as “online disinhibition effect” [4]. The range of his victims can increase considerably creating a larger group of affected individuals. Cyberbullying can produce psychological and emotional repercussion for the victim which includes lowering of self-esteem and self-confidence, depression, loneliness, anger, sadness, stress, degradation of health, and academic achievement and in the worst cases, self-injury and suicidal tendencies [5]. The experience of being a cyberbully has been frequently linked to psychological impacts that go all the way to the bully’s childhood years and external difficulties. Cyberbullies lack personal awareness, have low self-esteem, feel the need to have power over someone, portray depressive symptoms, have low self-efficacy, low empathy level, paranoia, phobic anxiety, and poor psychological well-being sometimes leading to suicide ideation.

Many people try to change themselves physically and mentally to be accepted into this fake world created by social media. That is, with the evolution of technology and uninhibited access to any information worldwide, the rate of cyberbullying has increased manifold. Cyberbullying Research Center in 2016 had published a study [6] which highlights that around 33.8% of the students between the ages of 12–17 years were prime victims of cyberbullying. Conversely, 11.5% of students between the age group of 12–17 accepted that they had been involved in cyberbullying. A study by McAfee, found that 87% of teens have witnessed cyberbullying.

In light of cyberbullying concerns, there is an urgent need for the detection and prevention of cyberbullying. The recent trend of shifting from reading to watching videos has led to a large diversity in the ages of people viewing YouTube, demanding that greater attention should be paid to prevent cyberbullying on this social media giant in comparison to others. Currently, social media platforms rely only on users alerting network software to remove the comments of bullies. The performance of alerting software can be improved by detecting instances and then removing the comments automatically. In this paper, four supervised machine learning [7] algorithms, namely random forest, sequential minimal optimization, k-nearest neighbor, and Naïve Bayes are implemented in "Waikato Environment for Knowledge Analysis" tool (Weka) to detect instances of cyberbullying across YouTube related to aspects like race, culture, gender, and physical attributes. The corpus for analysis comprises of comments gathered from arbitrary YouTube channels. The data is preprocessed and the results are evaluated based on measure like TPR, TNR, precision, recall, F-measure, and accuracy [8] for each selected algorithm. The following sections briefly represent the related work done in this area followed by a discussion of the proposed approach.

2 Background Work

Previously, a tremendous amount of work has been done for the detection of cyberbullying across various social media sites with Twitter being the most popular choice among researchers. A survey of existing research based on cyberbullying detection on well-liked social media giants like Twitter, YouTube, and Instagram have been done and tabulated below in Table 1.

From this table, we can infer that the most well-known social media sites are Twitter, YouTube, and Instagram when examined for detection of cyberbullying. As YouTube is the world's most notable user-generated content website with a viewership comprising almost the entire global Internet population, it is quite vulnerable to bullying through videos, the comments following, and most importantly, the replies to those comments. With about a 100 hours of videos uploaded every minute, the social networking site's acclaim, anonymity and almost negligible publication barriers allow users to upload content profane in nature. Thus, we can say that YouTube is highly susceptible to cyberbullying.

3 Proposed Approach

In this paper, we investigate the comment-based features to detect instances of cyberbullying in YouTube. Most of the researches in this area involve comments as a single attribute. However, the source of such profanity might be a part of some conversation or a response to a comment. Therefore, we consider the entire discussion or the reply comments as instances and manually label them as positive or negative instances of cyber bullying.

We extract the dataset for statistical machine learning using the social networking site's API for comments and their following conversations manually. Initially, we scraped comments from random YouTube channels. A manual inspection of the

Table 1 Survey of existing research in detecting cyberbullying

Social media	Author	Year	Journal/ Publisher	Technique
Twitter	Singh et al. [9]	2016	IEEE/ACM	Naïve Bayes
	Zhang et al. [10]	2016	IEEE	Pronunciation-based convolutional neural network (PCNN)
	Huang et al. [11]	2014	ACM	Bagging, J48, sequential minimal optimization, naive Bayes
	Santos et al. [12]	2014	Springer	Random Forest, J48 (decision tree), k-nearest neighbor, sequential minimal Optimization, naive Bayes
	Mangaonkar et al. [13]	2015	IEEE	naive Bayes, support vector machine, logistic regression
	Dadvar et al. [14]	2012	ACM	Support vector machine
YouTube	Dadvar et al. [15]	2012	ACM	Support vector machine
	Dinakar et al [16]	2011	AAAI	Naive Bayes, Rule based J-Rip, tree based J48, Support vector machine
	Dadvar et al. [17]	2014	Springer	Naive Bayes+ multi-criteria evaluation system, decision tree+ multi-criteria evaluation system, support vector machine+ multi-criteria evaluation system
	Dadvar et al. [18]	2013	BNAIC	Rule-based approach
Instagram	Miller et al. [19]	2016	IJCAI	Convolutional neural network (CNN) for clustering of Images
	Mattson et al. [20]	2015	University of Colorado	Naive Bayes, linear support vector machine

channel's comments showed that cyberbullying instances snowballed in case of controversial YouTube videos involving sensitive subjects like race (attacks on racial minorities like African American), culture (comments on the stereotypes attached to cultural traditions like Islam, Jewish), gender, sexuality (comments on LGBT Community), and physical attributes (topics related to the inherent characteristics of an individual).

Thus, for the detection of cyberbullying, YouTube videos on such sensitive topics are taken into account. The video IDs of the offensive videos acquired are annotated according to the type of offense. The video IDs generated and the respective offense category under which they fall has been listed in Table 2.

Table 2 Sample snippet of offensive videos

Video ID	Category	Comments	Rating	Date
2O7IW1CP0jA	LGBT	21,474	4.1	10/08/10
FrbC0ZGZ_pA	Culture	865	4.2	17/05/14
gQp0UugrpTk	Race	2942	4.1	20/11/15
KMU8TwC652 M&	Race	31,549	3.7	24/10/16
qXHPh3ecZEI	Physical attribute	1772	4	14/10/16

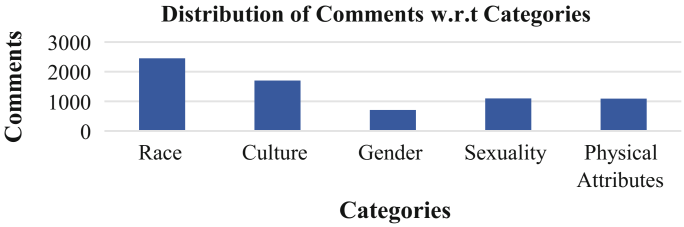

Fig. 1 Distribution of number of comments in each category

In totality, 7962 comments are scraped and labeled from 60 videos, which are roughly around 116 comments per video. A distribution representing the number of comments procured per category has been portrayed in Fig. 1.

The data obtained is preprocessed. The first step of preprocessing involves the correction of spelling mistakes in the comments. People tend to make short abbreviations for words such as "he's fy9" was converted to "He's fine". All the uppercase letters of comments are converted to lowercase characters. All the URL content, unwanted spaces and symbols such as "@$%#" are removed. Stopwords are also removed. These are words that are basically considered unwanted such "he, she, want, for, etc." The words are also stemmed to their basic form such as "hunting", "hunted", etc., reduce to the stem "hunt". They are then tokenized by breaking words into small segments called tokens. The next step is to select the features of the experimentation. The features are TF-IDF [21] (Term Frequency–Inverse Document Frequency); the Ortony lexicon of words [21] denoting negative abstract number density of foul words; frequently occurring part of speech unigram; and bi-gram tags observed in the training set across the dataset.

Using these features, different weights are assigned to the words in comments. 80% of the compiled dataset obtained is utilized for training data and the rest 20% as testing data. After being subjected to preprocessing to clean and organize the data, small datasets were evaluated using machine learning techniques, namely Random forest, Sequential Minimal Optimization (SMO), k-nearest neighbor, and Naïve Bayes.

4 Evaluation and Results

The results were analyzed using measures like TPR, TNR, precision, recall, F-measure, and accuracy. The terms negative and positive indicates the classifier's prediction while the terms false and true indicates whether that prediction corresponds to the user's judgment. 20-fold cross validation is used All these measures are calculated and portrayed in Table 3.

The results show that the best accuracy of 83% is given by KNN. Random Forest, Naïve Bayes, and Sequential Minimal Optimization closely follows KNN in decreasing order of accuracy with Random Forest giving an accuracy of 82.1%, Naïve Bayes of 81.4%, and SMO of 80.8%. The higher Precision and Recall Value indicates that a few false positives were detected. The graph in Fig. 2 compares the accuracy, precision, F-measure, and recall for all the training classifiers with the obtained results plotted on the y-axis and the selected classifiers on the x-axis.

According to the results, we deduce that among the above-applied algorithms, KNN has produced the highest accuracy of 83% when applied to cyberbullying detection of YouTube videos.

Table 3 Results obtained from testing the test dataset

Classifier	TPR	TNR	Precision	Recall	F-measure	Accuracy
KNN (k = 1)	0.830	0.308	0.824	0.830	0.826	0.830
Random Forest	0.822	0.324	0.815	0.822	0.817	0.821
SMO	0.809	0.389	0.799	0.809	0.797	0.808
Naïve Bayes	0.814	0.329	0.808	0.814	0.810	0.814

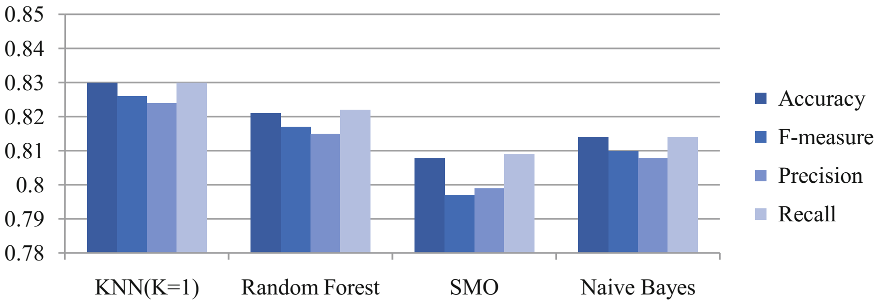

Fig. 2 Comparison of different measures for each of the training classifiers

5 Conclusion and Future Scope

This study empirically contrasted the analysis of cyberbullying detection on YouTube videos using four supervised machine learning algorithms, namely k-Nearest Neighbor, Random Forests, Sequential Minimal Optimization, and Naïve Bayes. 7962 comments were scraped and labeled from around 60 videos on YouTube, which are approximately 116 comments per video. These comments were based on sensitive topics like race, culture, gender, sexuality, and physical attributes which were then analyzed and the results were evaluated based on the various performances. The best accuracy is achieved using k-Nearest Neighbor, followed by Random Forest, Naïve Bayes, and Sequential Minimal Optimization. The experiments can be modified to include the use of comments and social networking graphs for better modeling of the problem. Various soft computing and deep learning techniques can also be used for cyberbullying detection.

References

1. Kumar A, Sebastian TM (2012) Sentiment analysis: a perspective on its past, present and future. IJISA 4:1–14
2. Agatston P-W, Kowalski R, Limber S (2007) Student perspective of cyberbullying. J Adolesc Health. Official publication of the Soc Adolesc Med 41:59–S60
3. Colette L (2013) Cyberbullying, associated harm and the criminal law. PhD Thesis, University of South Australia, pp 1–352
4. Suler J (2004) The online disinhibition effect. Cyberpsychology Behav 7:321–326
5. Foody M, Samara M, Carlbring P (2015) A review of cyberbullying and suggestions for online psychological therapy. Internet Inventions Elsevier 2:235–242
6. Robers S, Kemp J, Truman J, Synder T-D (2013) Indicators of school crime and safety: 2012. Bur Justice Stat 1–211
7. Bhatia MPS, Khalid AK (2008) Information retrieval and machine learning: supporting technologies for web mining research and practice. Webology 5
8. Bhatia MPS, Kumar A (2008) A primer on the web information retrieval paradigm. J Theor Appl Inf Technol 4(7):657–662
9. Huang Q, Singh V-K, Atrey PK (2014) Cyberbullying detection using probabilistic socio-textual information fusion. In: ACM international conference on advances in social networks analysis and mining (ASONAM), pp 3–6
10. Zhang X, Tong J, Vishwamitra N, Whittaker E, Mazer J-P, Kowalski R, Hu H, Luo F, Macbeth J, Dillon E (2016) Cyberbullying detection with a pronunciation based convolutional neural network. In: 15th IEEE international conference on machine learning and applications (ICMLA), pp 740–745
11. Huang Q, Singh V-K, Atrey P-K (2014) Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd international workshop on socially-aware multimedia ACM
12. Santos I, Miñambres-Marcos I, Laorden C, Galán-García P, Santamaría-Ibirika A, Bringas PG (2013) Twitter content-based spam filtering. In: Herrero Á et al (eds) International joint conference SOCO'13-CISIS'13-ICEUTE'13. Advances in intelligent systems and computing, vol 239, pp 449–458. Springer
13. Mangaonkar A, Hayrapetian A, Raje R (2015) Collaborative detection of cyberbullying behavior in Twitter. In: 2015 IEEE international conference on electro/information technology (EIT), pp 611–616

14. Dadvar M, Jong F (2012) Cyberbullying detection: a step toward a safer internet yard. In: Proceedings of the 21st international conference on World Wide Web ACM, pp 121–126
15. Dinakar K, Jones B, Havasi C, Lieberman H, Picard R (2012) Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans Interact Intell Syst 2
16. Karthik D, Roi R, Henry L (2011) Modeling the detection of textual cyber bullying. In: Cyberbullying social mobile web workshop at 5th international AAAI conference on weblog and social media, pp 11–17
17. Dadvar M, Trieschnigg D, de Jong F (2014) Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Canadian conference on artificial intelligence. Springer
18. Dadvar M, de Jong F, Trieschnigg D (2014) Expert knowledge for automatic detection of bullies in social networks. In: Canadian conference on artificial intelligence, pp 275–281
19. Zhong H, Li H, Squicciarini A-C, Rajtmajer S-M, Griffin C, Miller D-J, Caragea C (2016) Content-driven detection of cyberbullying on the instagram social network. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI-16), pp 3952–3958
20. Hosseinmardi H, Mattson S-A, Rafiq R-I, Han R, Lv Q, Mishra S (2015) Detection of cyberbullying incidents on the instagram social network. Technical report by Association for the advancement of artificial intelligence, pp 1–9
21. Ortony A, Clore GL, Foss MA (1987) The referential structure of the affective lexicon. Cogn Sci 11:341–364