International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 3, April 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

# Comparative Study of Cyberbullying Detection using Different Machine Learning Algorithms

**Rohini K R, Sreehari T Anil, Sreejith P M, Yedumohan P M**

Student, Computer Science and Engineering, Jyothi Engineering College (of KTU), Thrissur, Kerala, India

**ABSTRACT**

The advancement of social media plays an important role in increasing the population of youngsters on the web. And it has become the biggest medium of expressing one's thoughts and emotions. Recent studies report that cyberbullying constitutes a growing problem among youngsters on the web. These kinds of attacks have a major influence on the current generation's personal and social life because youngsters are ready to adopt online life instead of a real one, which leads them into an imaginary world. So, we are proposing a system for early detection of cyberbullying on the web and comparing different machine learning Algorithms to obtain the optimal result.

We are comparing four different algorithms which can be effectively used for the detection of cyberbullying, with the implementation of the bag of words algorithm with different n-gram methods. Comparatively naïve Bayes algorithm has the highest accuracy of 79% with trigram implementation of the bag of words algorithm.

***KEYWORDS:*** *cyber bullying, machine learning, naïve Bayes algorithm, decision tree algorithm, logistic regression algorithm, support vector machine algorithm*

***How to cite this paper:*** Rohini K R | Sreehari T Anil | Sreejith P M | Yedumohan P M "Comparative Study of Cyberbullying Detection using Different Machine Learning Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-3, April 2020, pp.1044-1048, URL: www.ijtsrd.com/papers/ijtsrd30765.pdf

IJTSRD30765



## 1. INTRODUCTION

The advancement of social media has an important role in the extended population of youngsters on the web. They utilize such platforms mainly for communicating and entertainment. The visible trend nowadays is communicating sentiments through social media. Most social media profiles look like a portrait of that user's life. The tendency of sharing every second of life through different forms of social networks has grown, where Instagram holds the top rank within youngsters. Hence this paper has chosen the Instagram platform and the necessary data collection is easier than others. The text-based analysis method we used for this research is facilitated by the availability of numerous public accounts and public comments related to them.

When we consider social media into account there are many safety issues, includes bullying, grooming, phishing etc. The aftereffect of this kind of issues is a huge area where it may lead to social, mental and physical issues in the current generation. In this research, we are mainly focusing on detecting bullying in social networks because the suicidal tendency in youngsters is an increasing issue in currents scenario. This kind of people expresses their feelings either as extreme depression or extreme anger which we will be able to identify through their posts, by considering sentiment in the captions, hashtags used in the posts etc. Different studies show that India has the highest occurrence of bullying through social media, such that it is necessary to control it.

Nowadays there are a lot of researches occurs related to bullying detection, avoidance etc. And the preliminary studies showing that the bullies targeting people who post something which is closely related to religious activities, sexual exposing, political activities and so on. Hence the first step which we did is that detection of this kind of posts from the Instagram network using some typical keywords related to this kind of posts and we extracted the meta-information and the comments too.

In this paper, we are comparing four different machine algorithms which can be used for detection bullying from comments and its related label with it. Using Natural Language Toolkit library in python we are implementing the naive Bayes algorithm, decision tree algorithm, logistic regression and support vector machine and then comparing their performance in terms of accuracy, precision and recall and we found that comparatively, the naive Bayes has a higher performance of 79% accuracy.

## 2. RELATED WORKS

There are a lot of researches which is already completed in this field. When we take them as a single one, we can find that Most of the existing systems use the SVM algorithm for the classification which at best provides an accuracy rate of 73%-76%. And the other things which are common in all papers are that they are based on an individual post and there is no history-based analysis. Also, there is no scope of

considering many features instead of it they are considering comments and their label. More than the papers when we consider the Instagram platform as a source, the current network giving many advantages over this kind of attacks to prevent them, but still there are many problems with it like they are giving warning or users who are searching things like self-harm, suicidal etc. which are the keywords closely related with this kind of issues, not only that the Instagram network introduced new stickers which can be used to say things like stop bullying, don't bully it etc. And most of the existing system which includes Content warning feature and Parental guide to monitoring the activities of the logged-in user. But there are still some problems with this kind of systems, for example when we consider the feature of the Instagram platform, we can find that there is jo warning for posts which includes the keywords mentioned above or there are no warnings for comments which seem to be bullying.

[1] F. Toriumi, T. Nakanishi, M. and K. Eguchi describes the clear difference between cyberbullying and cyber aggression in terms of frequency, negativity and imbalance of power applied in large scale labelling. Also uses images and their corresponding comments for the system. This was a multi-model classification result for cyberbullying detection. The main features they considered for the analysis are the content of the image, comments and metadata of the profile, and they found that cyberbullying occurs in posts which involves religion, death appearance and sexual hints. Other findings include, posts which face these attacks are most likely to be of negative emotions and relates to drug, tattoo etc. Here they use Latent Semantic Analysis (LSA) based on Singular-Value Decomposition (SVD) and linear support vector machine (SVM) classifier which uses n-gram texts with normalization. This system provides the highest of 79% Recall and 71% of Precision. But at the same time, the Data set is limited and they did labelling of data using survey, hence no media-based social networking can include in this system. Not only that there are no image reorganization algorithms are used, but the decisions are also taken only based on the survey.

[7] R. Badonnel, R. State, I. Chrisment and O. Festor, explains about the tracking of cyber predators in peer-to-peer network. This system mainly aims at detecting network attacks against vulnerable services and host-based attacks like unauthorized logins. Here they made the system capable of tracking and reporting cyber predators and hence it protects normal users from being in contact with these pathological users. Also defining the system using two criteria, tracking the deployment as well as tracking the target. And it is composed of a set of configurable honeypot agents and a central platform manager. The main managerial activities take place in this system are the generation of fake files, capturing of file requests and local statistical analysis. advantages of the system include fully compatible with management architecture and management protocol, fully independent of the file directory service and generic concern in the peer to peer clients. At the same time lack of central management and control over the available resources and the problem of central back up of files and folders make the system down.

When we consider the findings of [8] M. Di Capua, E. Di Nardo and A. Petrosino, they are tried to make a system which follows unsupervised learning methods for finding bullying activities in social media. This system proposes a method to detect cyberbullying with a hybrid set of features with classical textual features and social features. They adopt natural language processing algorithms and semantic as well as syntactic methods for filter the data. The main feature of this system is that they are considering emotional traces, and they make use of sentiment analysis with a set of features which are closely related to the social platform and the sentiment polarity of the sentences are calculated based on ranking. Here they are considered emojis and classified them as extremely negative, negative, neutral, positive and extremely positive. Here they use neural networks for the clustering purpose. The performance of the classifier is calculated based on precision, recall and F-measure.

Noviantho, S. M. Isa and L. Ashianti [9] explains how can we detect bullying using text mining techniques. Here the bullying conversations are identified based on naive Bayes method and SVM using a poly kernel. They used the data set from formsorng. me in the form of textual conversations and filtered it out through avoiding conversations which contains less than 15 words as well as which includes meaningless words. As an initial stage, they classify the data to two as Yes and No. Then a 4-class model development No, Low, Medium, High and an 11 class classification. Finally, they found that 4 class classification is the most optimal one and they proceed with that using n-gram five. This system employs based on textual conversations and hence they can only identify the cyberbullying behavior which is not enough for the system because of conversations contain other elements such as keywords and abbreviations and conversations include a lot of emoji contents but here it is not considered.

## 3. CHAL LENGES

- Imbalanced Class Distribution
- Issues related to cyberbullying definition
- Human data characteristics
- Culture Effect
- Language dynamics
- Prediction of cyberbullying severity
- Missing values
- Labelling of data
- Algorithm selection
- Performance measure selection

## 4. Methods and Results

We analyzed Instagram profiles and its associated posts for bullying detection. This section discusses data collection followed by filtering of relevant data, feature extraction, implementation and results.

### 4.1. Data Collection and Filtering

We are using Instagram data for analysis and the data was collected based on a set of keywords which includes Muslim, god, tattoo, pray, Hindu, bjp, etc. We chose these keywords by analyzing different profiles and found that these are the most commonly used keywords. The extracted fields of a profile from Instagram include the entire posts of the profile along with captions, hashtags, created time and the biography of the user and comments related with each post. The biography has an important role in ease filtering of data i.e., if bio contains words like quotes, memes, awareness, motivation etc. then we can avoid those profiles in the initial

stage. Also, these kinds of public pages will have more followers than followed by count. These points were used to filter the public pages like motivation pages or awareness pages etc. The we took the comments of each posts and labelled it manually as either bullying or non-bullying. A total of 1065 comments were considered and in which 636 were non bullying comments and the following are bullying one. We divide the entire data set as training as well as test data and each contains 746 and 319 respectively. During each algorithm calls for every n-gram combination we shuffle the dataset.

### 4.2. Feature Extraction

Here we consider each comment and initially preprocessed the data for the removal of stopping words, special characters etc. Then the features comments were vectorized or tokenized into different n-gram forms using the bag of words algorithms. Here first we tried unigram, bigram and trigram, then followed by the combination of these 3 in terms of the n-gram.

### 4.3. Algorithms Used and Results

Here we consider four machine learning algorithms for the comparative study and which are naive Bayes algorithm, decision tree, logistic regression and finally SVM. Here we use binary classification based on cross-validation of 5 members. We use nltk libraries for train the models and to compare their performance.

#### 4.3.1. Naïve Bayes Algorithm

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. The fundamental Naive Bayes assumption is that each feature makes an: independent and equal.

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .65 | .74 | .83 | .25 |
| Non-Bullying | .65 | .65 | .71 | .76 |

**Table1: Results of Naïve Bayes using unigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .68 | .77 | .77 | .35 |
| Non-Bullying | .68 | .67 | .67 | .79 |

**Table 2: Results using Bigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .79 | .78 | .44 | .56 |
| Non-Bullying | .79 | .71 | .91 | .79 |

**Table 3: Results using Trigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .75 | .78 | .14 | .24 |
| Non-Bullying | .75 | .63 | .97 | .76 |

**Table 4: Results of n-gram**

#### 4.3.2. Decision Tree Algorithm

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any Boolean function on discrete attributes using the decision tree. The parameters use in our system includes binary=True, entropy_cutoff=0.8, depth_cutoff=5, and support_cutoff=30.

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .64 | .57 | .61 | .42 |
| Non-Bullying | .64 | .68 | .72 | .73 |

**Table 5: Results of Decision Tree using unigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .66 | .56 | .56 | .52 |
| Non-Bullying | .66 | .7 | .7 | .74 |

**Table 6: Results of bigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .72 | .63 | .63 | .57 |
| Non-Bullying | .72 | .71 | .71 | .75 |

**Table 7: Results of Trigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .64 | .57 | .33 | .42 |
| Non-Bullying | .64 | .65 | .83 | .73 |

**Table 8: Results of ngram**

#### 4.3.3. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression) Parameters which we use includes algorithm='gis', trace=0, max_iter=10, and min_lldelta=0.5.

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .6 | .57 | .57 | .42 |
| Non-Bullying | .6 | .61 | .61 | .75 |

**Table 9: Results of Logistic Regression using unigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .62 | .57 | .57 | .52 |
| Non-Bullying | .62 | .62 | .62 | .76 |

**Table 10: Results of bigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .64 | .65 | .65 | .61 |
| Non-Bullying | .64 | .61 | .61 | .75 |

**Table 11: Results of trigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .73 | .57 | .33 | .42 |
| Non-Bullying | .73 | .6 | 1 | .75 |

**Table 12: Results of ngram**

#### 4.3.4. Support Vector Machine

Support Vector Machine is a discriminative classifier formally defined by a separating hyper plane. Here the given labelled training data uses the algorithm to give the optimal hyper plane which can classify new data []. An SVM model is the representation of data as points in space mapped so that the examples of the separate categories are divided by a clear gap that is wide as possible. In addition to this SVM's can efficiently perform a non-linear classification, implicitly mapping their inputs into high dimensional feature space.

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .64 | .58 | .58 | .43 |
| Non-Bullying | .64 | .6 | .6 | .75 |

**Table13: Results SVM using unigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .61 | .57 | .57 | .52 |
| Non-Bullying | .61 | .61 | .61 | .76 |

**Table14: Results of bigram**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .69 | .65 | .65 | .61 |
| Non-Bullying | .69 | .59 | .59 | .74 |

**Table15: Results of trigram.**

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Bullying | .76 | .58 | .34 | .43 |
| Non-Bullying | .76 | .6 | .95 | .75 |

**Table16: Results of ngram.**

## 5. CONCLUSION AND FUTURE WORKS

In this paper we are focused on the detection of cyberbullying using the Instagram data set, with help of four different machine learning algorithms and natural language processing algorithms like bag of words. And we found that comparatively naïve Bayes algorithms with trigram have the highest performance of 79% accuracy and other performance measures.

As a continuation we will Use this information to built a model for detecting bullying as well as cyber bullies and will monitor their activities. Also, we will try to implement a warning system for victims as well as bullies in current social networking sites. We will enhance our study to get more performance in terms of data set classifiers, NLP algorithms used and so on.

## 6. REFERENCES


[1] Hosseinmardi, Homa et al. "Prediction of Cyberbullying Incidents on the Instagram Social Network." ArXiv abs/1508.06257 (2015):

[2] M. Rybnicek, R. Poisel and S. Tjoa, "Facebook Watchdog: A Research Agenda for Detecting Online Grooming and Bullying Activities," 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, 2013, pp. 2854- 2859.

[3] V. S. Chavan and Shylaja S S, "Machine learning approach for detection of cyber- aggressive comments by peers on social media network," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, 2015, pp. 2354-2358.

[4] F. Toriumi, T. Nakanishi, M. Tashiro and K. Eguchi, "Analysis of User Behavior on Private Chat System," 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, 2015, pp. 1-4.

[5] R. Sugandhi, A. Pande, S. Chawla, A. Agrawal and H Bhagat, "Methods for detection of cyberbullying: A survey," 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), Marrakech, 2015, pp. 173-177.

[6] A. Upadhyay, A. Chaudhari, Arunesh, S. Ghale and S. S. Pawar, "Detection and prevention measures for cyberbullying and online grooming," 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2017, pp. 1-4

[7] R. Badonnel, R. State, I. Chrisment and O. Festor, "A Management Platform for Tracking Cyber Predators in Peer-to-Peer Networks," Second International Conference on Internet Monitoring and Protection (ICIMP 2007), San Jose, CA, 2007, pp. 11-11

[8] M. D Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyber bullying detection in social networks," 2016 23rd International Conference on Pattern Recognition (ICPR),Cancun, 2016, pp. 432-437.

[9] Noviantho, S. M. Isa and L. Ashianti,"Cyberbullying classification using text mining," 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Semarang, 2017, pp. 241-246.

[10] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flam-mini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. InFifth international AAAI conference on weblogs and social media, 2011

[11] Narendra M Shekokar and Krishna B Kansara. Security against sybil attack in social net-work. In2016 International Conference on Information

Communication and Embedded Systems (ICICES), pages 1–5. IEEE, 2016

[12] V S Subrahmanian and Srijan Kumar. Predicting human behavior: The next frontiers. Science, 355(6324):489–489, 2017

[13] George R.S. Weir, Fergus Toolan, and Duncan Smeed. The threats of social networking: Old wine in bottles? Information Security Technical Report, 16(2):38 – 43, 2011.Social Networking Threats.

[14] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. An-alyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter. In Proceedings of the 21st International Conference on World Wide Web, WWW '12, pages 71–80, New York, NY, USA, 2012. ACM.

[15] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter net-work. First Monday,15(1), 2010