IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Use Of Several Machine Learning Algorithms For Effective Prediction Of Cyberbullying Communication In Online Social Networks

I. Sree Geeta [*1], Umamaheswararao Mogili [*2]

[*1]M. Tech Scholar, Department of Computer Science & Engineering, *[2]Assistant Professor, Department of Computer Science & Engineering, St. Theressa Institute of Engineering and Technology, (Affiliated to Jawaharlal Nehru Technological University, Kakinada) ,Garividi ,Cheepurupalli, Vizianagaram Dist. – 535101.

ABSTRACT

In current days there are a lot of users who show interest in online social networks for communication especially in twitter, facebook and other networks. Although the usage has increased tremendously, there is still one limitation such as it is not providing security from bullying messages. The process of communicating with one another with the help of abused, vulgar, offensive messages in order to create fear for others in online platforms is known as cyberbullying. In a recent survey report almost 85 percent of messages contain bullying content and these messages are often posted by children, teenagers, pre-teenagers,above adults for communicating one with others on certain topics. There is no social media stopping these messages from communicating from one another on online platforms. Hence this motivated me to develop this proposed system in which cyber bullied messages are identified and removed from the communication. For this we try to take several machine learning algorithms and check the performance of each and every individual algorithm on the online social network communication dataset. By conducting various experiments on our proposed work by taking a sample dataset collected from KAGGLE, we finally came to the conclusion that Logistic Regression gives best accuracy compared with several models.

KEY WORDS:

Online Social Network, Cyberbulling, Machine Learning, Logistic Regression, Vulgar, Offensive.

1. INTRODUCTION

Online Social Media is defined as a group of Internet based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content. Via social media, people can enjoy enormous information, convenient communication experience and so on. However, social media may have some side effects such as cyber bullying, which may have negative impacts on the life of people, especially children and teenagers. Cyber bullying can be defined as aggressive, intentional actions performed by an individual or a group of people via digital communication methods such as sending messages and posting comments against a victim. Different from traditional bullying that usually occurs at school during face-to- face communication, cyber bullying on social media can take place anywhere at any time. For bullies, they are free to hurt their peers' feelings because they do not need to face someone and can hide behind the Internet. For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media. As reported in [2], cyber bullying victimization rate ranges from 10% to 40%.

In the United States, approximately 43% of teenagers were ever bullied on social media [3]. The same as traditional bullying, cyber bullying has negative, insidious and sweeping impacts on children [4], [5], [6]. The outcomes for victims under cyber bullying may even be tragic such as the occurrence of self-injurious behavior or suicides. One way to address the cyber bullying problem is to automatically detect and promptly report bullying messages so that proper measures can be taken to prevent possible tragedies. Previous works on computational studies of bullying have shown that natural language processing and machine learning are powerful tools to study bullying [7], [8]. Cyber bullying detection can be formulated as a supervised learning problem.

A classifier is first trained on a cyber bullying corpus labeled by humans, and the learned classifier is then used to recognize a bullying message. Three kinds of information including text, user demography, and social network features are often used in cyber bullying detection [9]. Since the text content is the most reliable, our work here focuses on text-based cyber bullying detection. In the text-based cyber bullying detection, the first and also critical step is the numerical representation learning for text messages. In fact, representation learning of text is extensively studied in text mining, information retrieval and natural language processing (NLP). Bag-of-words (BoW) model is one commonly used model that each dimension corresponds to a term. Latent Semantic Analysis (LSA) and topic models are another popular text representation models, which are both based on BoW models. By mapping text units into fixed-length vectors, the learned representation can be further processed for numerous language processing tasks. Therefore, the useful representation should discover the meaning behind text units.

2. LITERATURE SURVEY

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed Method. Literature survey is the most important step in software development process. For any software or application development, this step plays a very crucial role by determining the several factors like time, money, effort, lines of code and company strength. Once all these several factors are satisfied, then we need to determine which operating system and language used for developing the application. Once the programmers start building the application, they will first observe what are the pre-defined inventions that are done on same concept and then they will try to design the task in some innovated manner.

MOTIVATION

The main knowledge which is required to design this current application is text mining which is nothing but the process of dividing the text into multiple parts and where each and every individual part need to be saved in a separate array items.The term text mining, sometimes also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text data set. Normally from statistical pattern learning we mainly try to derive the patterns and trends that are calculated from the high quality of information. As we all know that the process of extracting or structuring the input text by identifying the main features and removing the unrelated data from that main document and finally convert the document in a structured way is known as text mining [10]. In this paper we use a word like high quality information, which is typically derived through a set of patterns and trends that are used in pattern learning applications. Also the term high quality in this paper clearly states that combination of some relevance and interestingness for the topic that was available in that conversation file. In the primitive text mining ,there are a possibilities like to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted

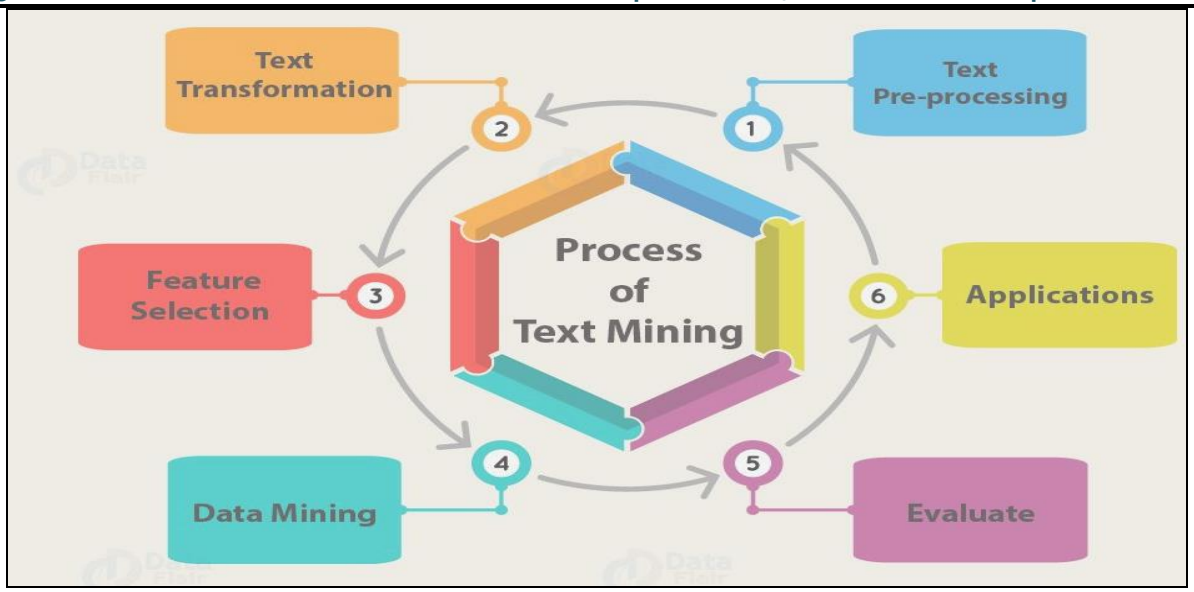


Figure. 1. Represents the Flow Diagram for Text Mining

From the above figure 1, we can clearly find out that it is a six layered approach where each and every individual layer is applied in a periodic manner to get the desired result. In the first layer we try to pre-process the text by applying some filters and now we need to transfer the text into some transformation manner. Once the text is transformed now we need to apply feature learning technique in order to identify the features which are matched with cyber bulled words and which are not matched with cyber words. Now here we need to apply data mining algorithms like K-means clustering in order to cluster the text into two categories one is those which are matched with cyber words and another is the set which is not matched with cyber list. Once the two lists are generated, the fifth level is used to evaluate the result and decide if there are any mistakes in the flow. If there are no mistakes, then the data can be send to the application level.

3. EXISTING METHODOLOGY

In the existing machine there has been no pre-described method or software to categorize the abused or cyber bulling messages for a text message which is posted on OSN walls and perceive the which meaning of that word or phrase and block that message not to be posted immediately on the customer's wall. So the following are the limitations that take place in the existing system. They are as follows:

LIMITATION OF EXISTING SYSTEM

1. Till now there was no approach in the literature. (To automatically find the cyber bullying messages and encode them proper right into a separate list.)
2. There was no class set of policies in literature. (that would robotically observe all the textual content that's posted by way of the usage of the customers and apprehend if there are any abused content material available on that posted messages.)

3. There isn't any term like BoW in the existing system, in which a bag of phrases is indexed right into a database and these bag of phrases are used for matching the dimensions of corresponding term that is published on the wall.

4. The primary problem of BoW is this may end up aware about the appropriate word in true message if the same message contains the word in plural way, this will not be diagnosed as matched word.

4. PROPOSED MODEL

Although tools are provided by contemporary social networking sites and laws are in place to fight cyberbullying, the majority of cyberbullying instances go unreported (Peterson, 2013). At the same time, there is no system in place for automatic detection of such behavior. Cyberbullying is one of the widely recognized problems which have a lasting impact on its victims. While healthy social behavior is the solution to this problem, social media platforms need to consider integrating tools and / or mechanisms that can help in the detection and prevention of such incidents. Therefore, to have a safer and more constructive social environment, it is necessary to design a smart network or an online patrol that will prohibit such behavior by monitoring and filtering the obscene, hateful, and improper content from social media posts.

DESCRIPTIONS OF DATASETS

Two datasets are used here. One is from Twitter/Facebook or any OSN user Dataset and another is from MySpace groups. The details of these two datasets are described below: Twitter Dataset: Twitter/Facebook is ‘’a real-time information network that connects you to the latest stories, ideas, opinions and news about what you find interesting ‘’ (https: //about.twitter.com/). Registered users can read and post tweets, which are defined as the messages posted on Twitter with a maximum length of 140 characters. The Twitter dataset is composed of tweets crawled by the public Twitter stream API through two steps. In Step 1, keywords starting with “bull” including “bully”,

”bullied” and” bullying” are used as queries in Twitter to preselect some tweets that potentially contain bullying contents. Retweets are removed by excluding tweets containing the acronym “RT”. In Step 2, the selected tweets are manually labeled as bullying trace or non-bullying trace based on the contents of the tweets. 7321 tweets are randomly sampled from the whole tweets collections from August 6, 2011 to August 31, 2011 and manually labeled 2. It should be pointed out here that labeling is based on bullying traces. A bullying trace is defined as the response of participants to their bullying experience. Bullying traces include not only messages about direct bullying attack, but also messages about reporting a bullying experience, revealing self as a victim et. al. Therefore, bullying traces far exceed the incidents of cyberbullying. Automatic detection of bullying traces is valuable for cyberbullying research. To preprocess these tweets, a tokenizer is applied without any stemming or stop word removal operations. In addition, some special characters including user mentions, URLS and so on are replaced by predefined characters, respectively.

Table 3.1: Statistical properties of two datasets

Statistics	Twitter	MySpace
Feature No	4413	4240
Sample No	7321	1539
Bullying Instances	2102	398

From table 3.1 represents the statistical of features are composed of unigrams and bigrams that should appear at least twice and the details of preprocessing can be found in.

MYSPACE DATASET

MySpace is another web2.0 social networking website. The registered accounts are allowed to view pictures, read chat and check other peoples' profile information. The MySpace dataset is crawled from MySpace groups. Each group consists of several posts by different users, which can be regarded as a conversation about one topic. Due to the interactive nature behind cyberbullying, each data sample is defined as a window of 10 consecutive posts and the windows are moved one post by one post so that we got multiple windows. Then, three people labeled the data for the existence of bullying content independently. To be objective, an instance is labeled as cyberbullying only if at least 2 out of 3 coders identify bullying content in the windows of posts. The raw text for these data, as XML files, have been kindly provided by Kontostathis et.al3 . The XML files contain information about the posts, such as post text, post data, and users' information, which are put into 11 packets.

P: Hi Hello How are you…..

B_P: I am fine … What about you.

P: I am good. Tomorrow we have class?

B_P: You are an idiot, stupid.

Figure: 2 MySpace Datasets

From the above Figure 2, we mainly focus on content-based mining, and hence, only extract and preprocess the posts' text. The preprocessing steps of the MySpace raw text include tokenization, deletion of punctuation and special characters. The unigrams and bigrams features are adopted here. The threshold for negligible low-frequency terms is set to 20, considering one post occurred in a long conversation will occur in at least ten windows. The details of this dataset is shown in Table 3.1.Since there were no standard splits of

training vs. test datasets in our adopted Twitter and MySpace corpora, need to define the training and testing datasets.

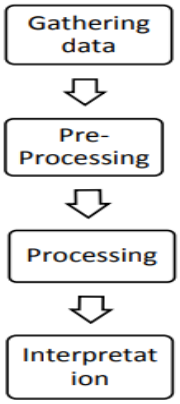
As analyzed above that the lack of labeled training corpus hinders the development of automatic cyberbullying detection, the sizes of training corpus are all controlled to be very small in our experiments. For Twitter dataset, randomly select 800 instances, which accounts for 12% of the whole corpus, as the training data and the rest data samples are used as testing data. To reduce variance, the process is repeated ten times so that can have ten sub-datasets from Twitter data. For MySpace dataset, also randomly pick 400 data samples as the training corpus and use the rest data for testing. The process is repeated ten times to generate ten sub data sets constructed from MySpace data. Finally, have twenty sub-datasets, in which ten datasets are from Twitter corpus and another ten datasets are from MySpace corpus.

5. PROPOSED MACHINE LEARNING ALGORITHMS

The proposed system contains ML Algorithms and we try to compare several ML classification algorithms in order to identify the cyber bulled messages and then try to predict the future occurrence based on word cloud. There are totally 4 modules present in this current application:

1) Gathering data,
2) Data Pre-Processing
3) Apply ML Models
4) Data Interpretation

The whole approach is depicted by the following flowchart.



1) DATA GATHERING

Here we try to load the data set from kaggle or UCI repository and once dataset is downloaded we try to load the dataset to the system for performing the operations.

2) DATA PRE-PROCESSING

Data preprocessing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Pre-processing for this approach takes 4 simple yet effective steps.

Training and Test data: Splitting the Dataset into Training set and Test Set Now the next step is to split our dataset into two. Training set and a Test set. We will train our machine learning models on our training set, i.e. our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to training set and therefore the remaining 20% to test set.

3) APPLY ML ALGORITHMS

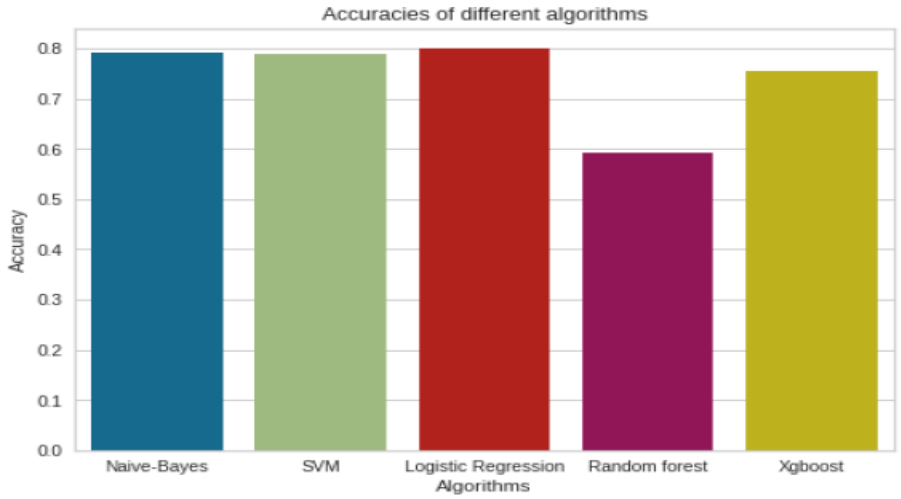
Once data is divided into test and train folders now we can apply well known ML Algorithms on the training data and then check the performance of each and every ML algorithm in order to predict the cyberbullying messages and then check which algorithm gives accurate and efficient result.

4) INTERPRETATION

The data set used for is further spitted into two sets consisting of two third as training set and one third as testing set. Here we apply several ML algorithms such as Naïve Bayes, SVM, Logistic Regression, Random Forest and Decision Trees to predict the cyberbullying and finally came to a conclusion that LR is best among all algorithms to predict the cyberbullying messages.

6. RESULT AND DISCUSSION

PERFORMANCE ANALYSIS



From the above window we can clearly see comparision of several ML algorithms and its accuracy for predicting the cyberbullying messages.This proposed application is getting more accuracy by using Logistic Regression and hence we finally conclude that LR is best accurate for cyberbullying prediction compared with other ML Models.

7. CONCLUSION

In this proposed article, we for the first time proposed and developed an application with several machine learning algorithms in order to detect cyberbullying messages and then predict the possibility of best algorithm in order to get accurate result by comparing several well known ML algorithms. By using this proposed approach, we can able to identify any abused a message which is send from one user to other user in an online social network. By conducting various experiments on our proposed framework, we finally came to a conclusion that our proposed approach is robust and easy to represent with word embedding technique. If this was applied in future for all types of OSN sites, we can tell that no OSN site will have any abused or vulgar contents in the communication and all the chat will be free from vulgar or offensive content.

8. REFERENCES

[1] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse processes, vol. 25, no. 2-3, pp. 259–284, 1998.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

[3] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[4] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[5] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying." in The Social Mobile Web, 2011.

[6] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," Communications in Information Science and Management Engineering, 2012.

[7] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in Advances in Information Retrieval. Springer, 2013, pp. 693–696.

[8] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

[9] P. Baldi,"Autoencoders, unsupervised learning, and deep architectures," Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7, p. 43, 2012.

[10] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," arXiv preprint arXiv: 1206.4683, 2012.