ISSN 2319-2720

Volume 7, No.2, April – June 2018

International Journal of Computing, Communications and Networking

Available Online at http://www.warse.org/ijccn/static/pdf/file/ijccn06722018.pdf

https://doi.org/10.30534/ijccn/2018/06722018

Identification of Online Abuse and It's Inhibition

Abhishek K M[1], Akshatha Kadaba [2], Bindu S [3],Lohithashree PS [4], Prof. Mangala C N[5]

[1]EWIT, India, abhiharsha18@gmail.com

[2]EWIT, India kadaba.akshatha @gmail.com

[3]EWIT, India, bindu.ys096@gmail.com

[4]EWIT, India, lohitha.arva @gmail.com

[5]EWIT, India, mangalacn@ewit.edu

ABSTRACT

Online abuse is an act of attacking an individual repeatedly with an intent to harm. This has a very disturbing effect on many individual irrespective of the age group. In this paper, we propose a new representation learning method to solve this problem. Our method named Naïve Bayes classifier is a simple probabilistic classifier based on Bayes' theorem with naïve independence assumptions between the features. Naive Bayes is an highly scalable with different number of parameters linear in number of predictors in a learning problem. The proposed method is able to exploit the hidden feature structure of abusive information and learn a robust and discriminative representation of text. We have implemented our algorithm using five lakhs of tweets and around one thousands of users.

Keywords: Naïve Bayes classifier, Bayes theorem, cyberbullying detection, Text mining.

1. INTRODUCTION

Now a day's social media being the greatest technology invented in today's world, a lot of people rely on it. There is a huge amount of exchange of information and other data among people. This creates a lot of problems for any individual. Any comments passed on an individual's post in a bad (negative) way might affect his/her mental state. A bad (negative) post might result in a disturbed mental condition of a person. As reported in [1], cyberbullying victimization rate ranges from 10% to 45%. In the United States, approximately 45% of teenagers were bullied on social media [2]. The same as traditional bullying, cyberbullying has negative, insidious and negative impacts on youngsters. [3], [4], and [5]. The outcomes for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behavior or suicides. This is a rising issue in the current technology trend affecting the youth's minds. A lot of proposals has been made in this case as to reduce the percentage of such cases on the social media data.

2. RELATED WORKS

In this section, we review several studies that used auto-encoder for detecting cyberbullying detection [9], [10], [15], [17].

We also discuss how our approach is different from these past studies.

Yin, Xue and Hong [4] use supervised learning, where each example is a pair consisting of an input vector and a desired supervisory signal.

A supervised learning algorithm analyses the pre - defined training data and produces an inferred function, which can be used for mapping new examples. Labelling using N-grams and weighting using TF-IDF.Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modelling. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Tf-idf is one of the most popular term-weighting schemes today; 83% of text-based recommend systems in digital libraries use tf-idf.

As per the stydu by Dinakar, Reichart and Lieberman [5] they collected youtube comments, labeled them manually and implemented various binary and multiclass classifications where it is the problem of classifying instances into one or more classes.[6] Reynolds used the decision tree and k-nearest neighbor (k = 1 and k = 3), labeling using

Amazon Mechanical Turk. It enables individuals and requesters to make the use of human intelligence to perform tasks that computers don't do Workers can then browse among existing jobs and complete them in exchange for a monetary payment set by the employer. For placing jobs, it uses an API, or the more limited MTurk Requester site.

We believe that our approach has an advantage over the strategies used in the past studies because ours is robust to the change in the features of detecting bullying words by accepting huge number of data.

2. METHODLOGY

The methodology used in this research is as follows;

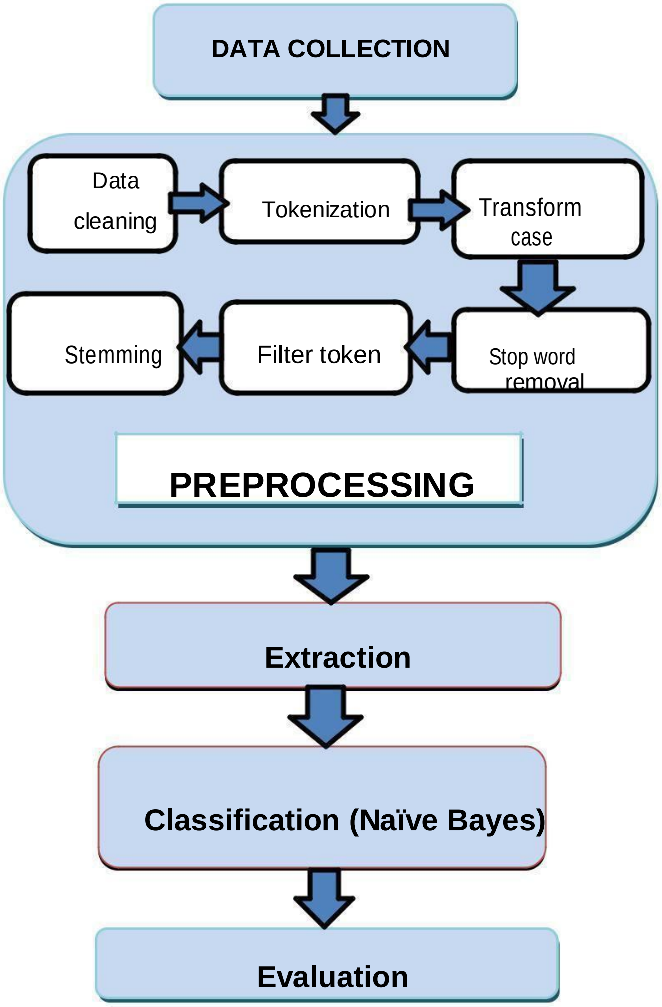


Figure 1

2.1 Data Collection

In order to build a cyberbullying classifier a manually labelled pre-defined dataset is required. A few labelled datasets are available, but it is recommended that you create and label your own dataset based upon the social media platform that you need to integrate with.

In this stage the classification used is Naïve Bayes. Each conversation in the form of comments are combined into one text conversation. The collected text conversations are randomly divided into sets of training data. Each text conversation consisting of 1600 conversations is labelled according to the data set and text conversation status.

2.2 Pre-processing

Pre-processing is an important task and critical step in Text mining, Information retrieval (IR) and Natural Language Processing (NLP). In the area of Text Mining, data pre-processing used for extracting interesting and non-trivial and knowledge from unstructured text data. Information Retrieval (IR) is a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information.

The user's need for information is represented by a query or profile, and contains one or more search terms, plus some additional information such as weight of the words. Hence, the retrieval decision is made by comparing the terms of the query with the index terms (important words or phrases) appearing in the document itself. The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to query.

2.2.1 Data Cleaning and balancing

Data cleaning is done with the help of Microsoft excel by eliminating conversations that have total characters under 15 letters, deleting meaningless words like "uhaha", "hehe", "kwkw", "emnn", "umm".

For the purposes of data balancing on the classification of 2 classes (cyberbully, non-cyberbully), 4 classes (non-cyberbully, cyberbully level severity low, cyberbully level severity middle, cyberbully level severity high), and 11 classes (non-cyberbully, cyberbully level severity 1
– 10), then the data used amounted to 1.600 for balancing data (800 labelled cyberbully and 800 labelled non-cyberbully)with the following allocation

a) 2 Class: each class amounts to 800 data

Class No is: 800 data with label severity 0
Class Yes is: 800 data with label severity 1-10.

b) 4 Class: each class amounts to 240 data

Class No: 240 data with label severity 0
Class Low: 240 data with label severity 1- 3

Class Middle: 240 data with label severity 4 -7
Class High: 240 data with label severity 8 -10.

c) 11 Class: each class amounts to 80 data

Class 0 : 80 data with label severity 0
Class 1 : 80 data with label severity 1
Class 2 : 80 data with label severity 2
Class 3 : 80 data with label severity 3
Class 4 : 80 data with label severity 4
Class 5 : 80 data with label severity 5
Class 6 : 80 data with label severity 6
Class 7 : 80 data with label severity 7
Class 8 : 80 data with label severity 8
Class 9 : 80 data with label severity 9
Class 10: 80 data with label severity 10.

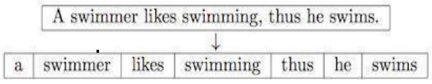
2.2.2 Tokenization

Tokenization is a method of breaking up a sequence of sentences into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining. Tokenization is used in computer science, where it plays a large part in the process of lexical analysis. Tokenization depends mostly on simple heuristics in order to separate tokens by following a few steps:

Tokens or words are separated by whitespace, punctuation marks or line breaks are not included.

All characters within contiguous sentences are part of the token. Tokens can be made up of all alpha characters, alphanumeric characters or numeric characters only.

Table I: Example for tokenization



A swimmer likes swimming, thus he swims.

↓

a	swimmer	likes	swimming	thus	he	swims

2.2.3 Stop Word Filtering

Stop-word filtering consists in eliminating stop-words, i.e., words which provide little or no information to the text analysis. Stopping is the process of reducing each word (i.e., token) to its stem or root form, by removing its suffix. The purpose of this step is to collection words with the same theme having closely related semantics**.**

2.2.4 Stem Filtering

Stemming is the method of reducing inflected words to their word stem, base or root form generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root Stem filtering consists in decreasing the number of stems of each Sentence. In specific, each sentence is filtered by eliminating from the set of stems the ones not going to the set of related stems. For example you can see below that *banks* and banking become *bank, and investing* and *invested* become *invest.*

The classifier doesn't understand that the verbs *investing* and *invested* are the same, and treats them as different words with different frequencies. By stemming them, it groups the frequencies of different inflection to just one term

— in this case, *invest.*

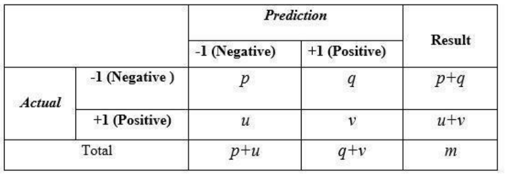
3 Extraction

The pre-processing text conversations will be transformed into a vector space model where text conversations are represented with a vector of extracted features. Features resulting from the extraction are words or combinations of words to form a list of words and the calculation of the weight with TF-IDF

4 Classification

In this stage the classification will use the Naïve Bayes & SVM method. Each conversation in the form of questions and answers is combined into one text conversation. The collected text conversations are randomly divided into sets of training and test data. Each text conversation consisting of 1600 conversations is labelled according to the data set and text conversation status. The division of text conversations into data sets is done 10 times.

Table II Confusion Matrix



		Prediction		Result
		-1 (Negative)	+1 (Positive)	
Actual	-1 (Negative)	*p*	*q*	*p+q*
	+1 (Positive)	*u*	*v*	*u+v*
Total		*p+u*	*q+v*	*m*

5 Evaluation

To evaluate the classification model based on the accuracy can be measured from the accuracy of the model with the method called confusion matrix. Confusion matrix is a matrix consisting of rows and columns as shown in Table II. The row corresponds to a predefined value while the column corresponds to the predicted value predefined by the classification model [8].

3. EXPERIMENTAL RESULT

This estimates the conditional probability of a particular word/term/token given a class. If the word is bullying word then the comment will be deleted or if the word is legitimate comment. Then the word is displayed on the screen with the accuracy of 98%.

4. CONCLUSION

This project solves the text-based online harassment recognition issue, where vigorous and selective depiction of tweets are crucial for an efficient recognition system. Being cruel to others by sending or posting harmful material using technological means, involving electronic technologies to facilitate deliberate and repeated harassment or threat to an individual or group. Repeatedly sending offensive messages, rude and insulting messages that include threats of harm or highly intimidating.

5. ACKNOWLEDGEMENT

We wish to offer our sincere gratitude to our principal Dr. K. Channakeshvalu for providing me an opportunity to be a part of the East West Institute of Technology, Bangalore under the graduate program and moreover to allow me work on this final year project on the respective domain.

We would like to offer gratitude to the H.O.D of Computer Science and Engg, East West Institute of Technology, Dr. Arun Biradar for his continuous support and encouragement towards our project.

We sincerely thank our guide Associate Proff. Mangala C N, for the guidance and encouragement in carrying out this work and rectifying the mistakes in the same for the betterment of our work.

REFERENCES

[1]. A. M. Kaplan and M. Haenlein, “Users of the world, unite! the challenges and opportunities of social media,” Business horizons, vol. 53, no. 1, pp. 59–68, 2010.
https://doi.org/10.1016/j.bushor.2009.09.003

[2]. R. M. Kowalski, G. W. Gamete, A. N. Schroeder, and M. R. Lattanner, “Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth.” 2014.

[3]. M. Ybarra, “Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression,” National Summit on Interpersonal Violence and Abuse across the Lifespan: Forging a Shared Agenda, 2010.

[4]. B. K. Biggs, J. M. Nelson, and M. L. Sampilo, “Peer relations in the anxiety–depression link: Test Of a mediation model,” Anxiety, Stress, & Coping, vol. 23, no. 4, pp. 431–447, 2010.
https://doi.org/10.1080/10615800903406543

[5]. C. Vercellis, Business intelligence: Data Mining and Optimization for Decision Making, Politecnico di Milano, Wiley, 2009.
https://doi.org/10.1002/9780470753866

[6]. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, “Bullying words from social media,” in Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational

Linguistics, 2012..

[7]. Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. ACM, 2014, pp. 3–6. 1949-3045 (c) 2015 IEEE.

[8]. 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.

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

[10]. V. Nahar, X. Li, and C. Pang, "An effective approachforcyberbullyingdetection," Communications in Information Science and Management Engineering, 2012.

[11]. H. Kwak, C. Lee, H. Park, and S. Moon, "twitter, a social network or what?" in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 591–600.

[12]. C. Vercellis, Business intelligence: Data Mining and Optimization for Decision Making, Politecnico di Milano, Italy.: Wiley, 2009. https://doi.org/10.1002/9780470753866

[13]. D. Yin, Z. Xue, & L. Hong, "Detection of Harassment on Web 2.0". Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009, 1-7.

[14]. A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter," in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007, pp. 56–65. https://doi.org/10.1145/1348549.1348556

[15]. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007, pp. 29–42. https://doi.org/10.1145/1298306.1298311

AUTHORS BIOGRAPHY

Mrs.Mangala.C.N received the B.E degree in Computer science and Engineering from NCET, Bangalore, VTU University in 2006 and got M.Tech degree in Computer Science from RVCE, Bangalore, India. She is currently working as Associate Professor in the faculty of CSE, EWIT-Bangalore, India. Her area of interest includes Image Processing, Data Mining and Big Data.

Mr.Abhishek K M is pursuing his 8th sem B.E in Computer Science and Engineering in East West institute of Technology, Bangalore, India. His area of interest includes Big Data and data mining.

Ms.Akshatha Kadaba is pursuing her 8th sem B.E in Computer Science and Engineering in East West institute of Technology, Bangalore, India. Her area of interest includes Big Data and data mining.

Ms.Bindu S is pursuing her 8th sem B.E in Computer Science and Engineering in East West institute of Technology, Bangalore, India. Her area of interest includes Big Data and data mining.

Ms.Lohithashree P S is pursuing her 8th sem B.E in Computer Science and Engineering in East West institute of Technology, Bangalore, India. Her area of interest includes Big Data and data mining.