Indian Journal of Science and Technology, *Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/93825, July 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter

K. Nalini[1] andL. Jaba Sheela[2]

[1]Bharathiyar University, Coimbatore - 641046, Tamil Nadu, India; immanuelsamen@rediffmail.com
[2]Panimalar Engineering College, Chennai - 600123, Tamil Nadu, India; sujitha14@hotmail.com

Abstract

Objectives: Social networks are becoming a risk for minors especially those are using it regularly. This action can also lead to Cyber bullying. The unstructured texts which are present in the enormous amount of information cannot simply be used for further processing by computers. So, the specific preprocessing methods and algorithms are needed in order to extract useful patterns. **Methods/Analysis**: One of the important research issues in the field of text mining is Text Classification. The Twitter corpus is used as the training and test data to build a sentiment classifier. The positive or negative sentiments of a new tweet are used to detect Cyber Bullying messages in Twitter using LDA with Naive Bayes classifier. **Findings:** The result shows that our model gives the better result of precision, recall and F-measure as nearly 70%. Naive Bayes is the most appropriate algorithm comparing with other algorithms like J48 and Knn. The CPU processing time for Naive Bayes algorithm is comparatively less than the other two classification algorithm. **Improvements:** The performance of the system can be improved by adding extra features to more amount of data.

Keywords: Cyber Bullying, LDA, Naive Bayes, Text Mining, Twitter

1. Introduction

The modifications and transformations of relationships and communication methods put bullying behavior into a new format commonly referred to as Cyber bullying. Many teenagers from countries have exposed about the life-threatening bullying experiences. So, there is a necessary to draw special care to it. Bullying has occurred in various forms of confusions in the social network. One form of online misbehavior which has deeply affected society with harmful consequences is known as Cyber bullying. Traditional bullying used to be a demonstration of dominance and consolidation of social status by making use of physical power and creating fear and discomfort for those who were weaker and vulnerable. Cyber bullying is described as a deliberate act that is conducted through digital technology to hurt someone. The proposed method aims to accurately detect harmful messages and twitter data has been used for sentiment analysis. First, the key terms are identified using the Latent Dirichlet Allocation (LDA). Each tweet in the n-dimensional vector is represented by these key terms. In order to find the sentiment of each tweet, we build a sentiment classifier, by using tweet vectors. The result of the experiment shows that our proposed method is efficient and effective. The main aim of this paper is to use sentiment analysis to detect bullying instances in Twitter.

1.1 Related Work

Latent Dirichlet Allocation is a flexible generative probabilistic model[1] for a collection of discrete data and it can be readily embedded in a more complex model. In a recent study, the principal component analysis is used for the feature reduction and feature selection[2] for sentiment analysis using decision forest method. In another study[3] rule-based approach is used in chat log data set to detect

**Author for correspondence*

Cyber bullying. In other intriguing works[4] the datasets of chat room were used to generate the local features and sentiment features. In a study of detecting[5] Cyber bullying the features of gender specific were used to categorize the male and female groups. The keyword search method[6] is used to detect the sexual predation in chat log data set to differentiate between predator and victim. In another study[7], the count and normalization of the bad words are used to assign the severity level of the bad words list in the website, Formspring.me. It[8]is considered, not reverent comments and sexual messages to detect the Cyber bullying in you tube.

1.2 Research Motivation

Cyber bullying is one of the problems which emerged with the growing use of social networks. Most of the teenagers and adolescents are active on social networks. Based on a recent annual Cyber bullying surveys conducted on teenagers and adolescents from the UK, the USA, Australia and other countries, 7 out of 10 young people have been the victim of Cyber bullying. The survey showed that the top three social networks frequently used by Internet users are Face Book (75%), YouTube (66%) and Twitter (43%). These three social networks are also found to be the most common networks for Cyber bullying as 54%, 21% and 28% of their users have experienced Cyber bullying respectively. Cyber bullying leads to suicidal thoughts and some of the youngsters who are bullied regularly by traditional bullying, likely to attempt suicide. There have been a few prominent cases throughout the world involving youngsters taking their own lives to some extent due to the harassment over the Web. In light of these studies and the suicide cases reported in broad communications, we proposed, as our work, a product base for deriving and envisioning harassing occasions in Twitter.

1.3 Aim

We aim to apply Text Mining techniques to social issues in our community on the Internet. All the more particularly, our major goal is to detect tormenting occasions in Twitter and build their permeability so that social organizations could make a move; e.g., legitimate direction to victims and bullies.

1.4 Concept

Text mining in Twitter is a new but interesting.

These are few reasons for using Twitter data for the sentiment analysis.

- People can express and share their thoughts and ideas about various titles in Twitter.
- It also contains a large number of tweets and it increases every day.
- There are different types of users like cinema stars, politicians and ministers from different countries on Twitter. So, there is a possibility of collecting different types of tweets from different types of users.

We have collected five thousand tweets and they are distributed as two sets of sentiment as:

1. Tweets contain non-bullying words as positive sentiment.
2. Tweets contain bullying words as a negative sentiment.

Sentiment analysis is an extraordinary instance of text mining, for the most part, centered on recognizing opinion polarity, keeping in mind it's frequently not very accurate, it can at present be helpful as the premise for identifying harassing instances in Twitter. Since our main objective is to detect bullying instances, we will concentrate just on the negative sentiment tweets. The corpus is divided into training data and test data, in order to build a sentiment classifier using LDA methods. These classifiers are used to find the positive tweets, negative tweets and neutral tweets. The different sections of the paper as follows, Section 2 depicts Research Methodology. The Results of the Experiments is given in Section 3 and Section 4 contains Conclusion.

2. Research Methodology

The complete structure of our model is described in Figure 1 in order to arrange the tweets depending on their sentiments.

The following passages explain the functionalities depicted in Figure 1.

2.1 Tweets Slithering

The search key is used in Twitter crawling to download the tweets from the Twitter database. The Twitter's Application Programming Interface "twitterAj-core-4.02.jar" is utilized for this purpose. The users' connected data

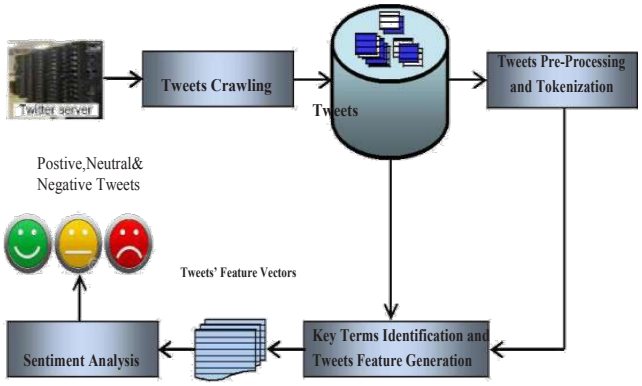

Figure 1. System architecture

and information regarding tweets are obtained by this API using its classes and techniques. We can also fetch area and speech based tweets using this API. Based on our requirements the fetched tweets can be saved in database or text file.

2.2 Tweet Pre-Processing and Tokenization

The unwanted tokens are filtered from the tweets by tweet Preprocessing and Tokenization. The words containing special symbols, stop words, Retweets, mentions, URLs are filtered out from tweets. A bag of words is formed by splitting the remaining part of the tweets as tokens dependent upon clear space and punctuation mark.

2.3 To Identify Key Terms and Tweets' Characteristics Vector Creation

In an n-dimensional feature vectors, the identification of key terms and the feature generation of tweets' are focused on modeling the each tweet. Each token of a tweet acquired from the past methodology is recognized as a candidate term. The set of tweets is altered into a term-tweet grid A of order m x n. In this grid, a row denotes a candidate term and a column denotes a tweet. The basic part $a_{i,j}$ of grid A, persisted as the weight of term t_i in j^{th} tweet using tf-idf method, which is obtained using Equations 1 and 2.

$$a_{i,j} = tf(t_{i,j}) \times idf(t_i) \tag{1}$$

$$idf(t_i) = \log \frac{n}{\left|\{tw_j : ti \in tw\}\right|} + 1 \tag{2}$$

We apply Singular Value Decomposition (SVD) to connect feature set under a low dimensional area. This expands the efficiency of the recommended frame work both in terms of memory and processing time. For a given m x n matrix with m ≥ n, the SVD does partition under an m x n orthogonal matrix U, an n x n diagonal matrix S and an n x n orthogonal matrix V such that A = USV'. In this partition, U denotes the term matrix and V denotes the tweet matrix. Each row of matrix V denotes a tweet vector which is deducted from m to n in the new characteristics space. We clubbed tweets into a number of groups that is used to build the input file for LDA, based on matrix V.

We utilize Latent Dirichlet Allocation (LDA), to acquire the candidate term. The group of tweets is used to generate an input file for LDA. In this file, the first line consists of an integer value k denoting the number of clusters. Followed by this, there are k paragraphs; one for each cluster, containing the list of terms obtained from the corresponding tweets belongs to that cluster. To get Θ and Φ matrices, we have utilized JGibbLDA to execute LDA on the dataset and Dirichlet hyperparameters, α and β are assigned as 0.1 and 0.5, respectively. The components of the Φ matrix and the Θ matrix denote the term-topic and topic cluster distributions, respectively. The Θ and Φ matrices are used to assign a ranking score to each term using Equations 3 and 4. After evaluating the score of each term, we formed them in diminishing order of their scores and to find top n-terms as key terms.

$$score(t_i) = \max_{j=1}^{n} \{\Phi_{j,i} \times \omega_j\} \tag{3}$$

$$\omega j = \sum_{i=1}^{k} n_i \times \Theta_{i,j} \tag{4}$$

Based on the occurrence of the term (0 or 1), each tweet is designed as an n-dimensional double characteristics vector and they are used in training and testing of sentiment classifier.

2.4 Classification

The double characteristic vectors of the tweets are utilized as input for sentiment analysis. The Naive Bayes classifier depends on Bayes' theorem and it is utilized for classifying the tweets as a positive tweet, negative tweet or neutral based on the text. If S is the sentiment of a provided tweet T then the probability is determined by Equation 5.

$$P(S/T) = P(S) * P(T/S) / P(T) \tag{5}$$

3. Experimental setup and Results

The test setup and outcomes are introduced in this section. For the assessment of our model, we have used 3200 tweets, which are downloaded using Twitter's API. The facts about the downloaded tweets are displayed in Table 1. The positive sentiment, negative sentiment or neutral of each tweet are assumed by the intelligent people based on a message.

The very important task in this system is to identify the key terms. A numeric score is assigned to each word of the tweets by LDA and depending upon the score value they are arranged in descending order. Table 2 depicts the terms which in the top. The 6600 key terms are found as total key terms, after performing Tweets' preprocessing and tokenization of those 3200 tweets.

These key terms are used in order to generate the feature vectors of the tweets. To train and test the classifier, an input file containing top n key terms were utilized. The top 1000, 2000, 3000, 4000, 5000, 6000 and 6600 key terms are stored in input files to evaluate. The generation of the input file is done using a Java program and it reads the details of key terms. It also helps to find the number of tweets and also it create the input file depends on the sentiment of the tweet.

The major goal of this process is to analyze and to classify the given tweet text into non-bullying or bullying depending on the sentiment of the tweet. If a tweet is analyzed correctly and it is same based on the assignment of an expert, then we assure that it is correctly classified. The recall, precision and F-measure values are used in order to evaluate the system and they are explained below.

Precision (π): The ratio of true positives among all retrieved instances.

$$\pi = TP / (TP + FP) \tag{6}$$

Table 1. Tweets' data set statistics

Tweet Category		Tweets' Statistics				Users' Statistics	
	No. of tweets	Avg. no. of hash tags	Avg. no. of URLs	Avg. no. of mention	Avg. no. of Followers	Avg. No. of Friends	Avg. no. tweets
Non bullying	2000	1.9	0.37	0.95	2104.4	1093.84	18865.53
Bullying	1200	0.54	0.49	1.03	2352.48	600.97	29707.23
Grand Total	3200	0.94	0.48	0.95	2061.99	923.65	24130.86

Table 2. Key terms and their LDA score

Key Terms	LDA Score	Key Terms	LDA Score	Key Terms	LDA Score	Key Terms	LDA Score
Fuck	96.91	Suck	67.13	Lick	63.16	stupid	32.97
Ass	95.73	Ugly	65.46	hell	58.14	bastard	32.18
Shit	90.17	Naked	65.46	bitch	57.41	sucko	31.08
Bullshit	87.40	Sexy	67.13	Hotbitch	33.76	freak	30.59
Gay	84.57	Boo	63.99	sipper	32.97	fat	30.35
Dumb	72.77	Mood	63.26	Kill	32.18	dirty	29.79

Table 3. Evaluation of key terms

No. of Terms	TP Rate	FP Rate	Precision	Recall	F-Measure
1000	0.688	0.213	0.689	0.688	0.687
2000	0.707	0.207	0.705	0.706	0.704
3000	0.705	0.215	0.702	0.705	0.702
4000	0.702	0.215	0.698	0.702	0.699
5000	0.701	0.214	0.698	0.701	0.698
6000	0.708	0.218	0.703	0.708	0.703
6600	0.706	0.221	0.701	0.706	0.701

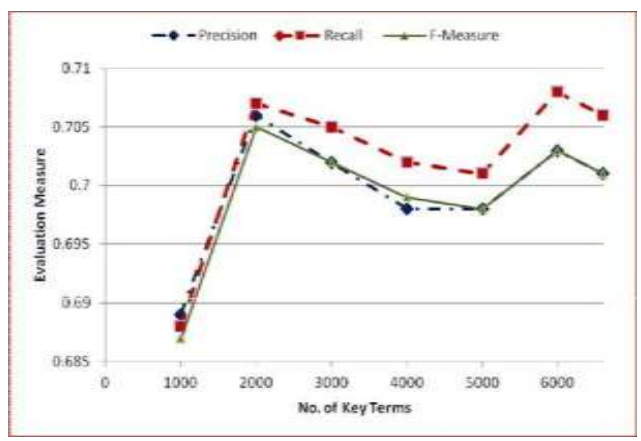

Figure 2. Precision, recall and F-measure for different values.

Table 4. Evaluation for the classification algorithms

Classification Algorithms	No. of Terms	TP Rate	FP Rate	Precision	Recall	F-Measure
J48	2000	0.678	0.203	0.679	0.668	0.677
Naive Bayes	2000	0.707	0.207	0.706	0.707	0.705
Knn	2000	0.700	0.204	0.668	0.700	0.688

Recall (ρ): the ratio of the positives among all positive instances.

$$\rho = TP / (TP + FN) \tag{7}$$

F-measure (F): the harmonic mean of recall and precision.

$$F = 2\rho\pi / (\rho + \pi) \tag{8}$$

The Naive Bayes classifier with 10 fold is used in order to classify a database consisting of a various count of key terms. The evaluation summary of the system listed in Table 3 and Figure 2 shows the respective graph. The table shows that our model gives better execution results if we consider one-third of the total key terms as feature attributes and it gives the best result when n is equal to two thousand key terms.

Table 4 exhibits the evaluation of classification algorithms in FPR, TPR, Precision, Recall and F-measure. When n = 2000 key terms, Naive Bayes shows the result as F-measure = 0.705 and it is the most appropriate algorithm comparing with other J48 and Knn. The CPU processing time for Naive Bayes algorithm is comparatively less than the other two classification algorithm.

4. Conclusion

The work exhibited here on the best way to go up against the marvel of Cyber bullying epitomizes the potentially included benefit of taking a multidisciplinary point of view. Cyber bullying is an old social wonder that is established in human instinct. Cyber bullying is a later variation led utilizing digital infrastructure. Sentiment analysis model is implemented in order to detect the Cyber bullying in Twitter and the tweets are classified as positive or negative. The key terms identification is the first step in this system. LDA method is used for that purpose and depending on LDA value the identified key terms are maintained in decreasing order. Then we created the feature vectors of each and every tweet by considering top n key terms as attributes. Each tweet is changed into a binary feature vector. Then, the system is trained by Naive Bayes classifier. The model gives the best outcome as 70.5% precision, 70.6% recall, and 70.4% F-measure by taking one-third of a total number of recognized key terms. In future, the performance of the system can be improved by adding more features with a large set of data.

5. References

1. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003 Mar; 3(4-5):993–1022.
2. Jeevanandam J, Koteeswaran S. Feature selection using random forest method for sentiment analysis. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–7.
3. Mcghee I, Bayzick J, Kontostathis A, Edwards L, Mcbride A, Jakubowski E. Learning to identify Internet sexual predation. International Journal on Electron Commerce. 2011 Apr; 15:103–22.
4. Yin D, Davison BD, Xue Z, Hong L, Kontostathis A, Edwards L. Detection of harassment on Web 2.0. Proceedings of the Content Analysis in the Web 2.0. (CAW2.0) Workshop at WWW2009; 2009 Apr.
5. Dadvar M, Jong FD, Ordelman R, Trieschnigg D. Improved Cyber bullying detection using gender information. Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012); 2012 Feb. p. 23–5.
6. Kontostathis A, Edwards L, Leatherman A. ChatCoder: Toward the tracking and categorization of Internet predators. Proceedings of Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining (SDM 2009); 2009 May.
7. Reynolds K, Kontostathis A, Edwards L. Using machine learning to detect Cyber bullying. Proceedings of the 2011 10th International Conference on Machine Learning and Applications Workshops (ICMLA 2011). 2011 Dec; 2:241–4.
8. Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual Cyber bullying. International Conference on Weblog and Social Media - Social Mobile Web Workshop; Barcelona, Spain. 2011.