An Efficient Feature Extraction Method For Mining Social Media

V Mageshwari, Dr. I. Laurence Aroquiaraj

Abstract: Social media facilitates the users to exchange their opinion, thoughts and ideas. The advantage of sharing an information through social media is, it will widespread the content quickly. There are so many social media platforms among which Twitter is one of them. Through twitter the user can communicate the information briefly. So many real-world issues are discussed on twitter, in which the discussion about HIV/AIDS is ranked as one of the topmost topics. Due to the advancement of social media many users have come forward to discuss about this societal topic. These kinds of discussion will help the communication campaigns to promote better HIV/AIDS education. In this work tweets were collected by the keywords including HIV and AIDS. Following the pre-processing steps, feature extraction has been carried out. Feature extraction is very crucial step in mining twitter because the data is in unstructured format. So, increasing the efficiency of feature extraction will improve the outcome of classification task. In this work an efficient feature extraction method has been proposed which gives a better result when compared to existing.

Index Terms: Classification, BOW, feature extraction, HIV/AIDS, pre-processing, tf-idf, twitter

1. INTRODUCTION

HIV/AIDS is remaining as a public health problem. There are so many communication campaigns making effort to minimize the spread of HIV. They are actively involved in creating prevention and awareness among people. The information shared on social media will help them more in understanding what people speak about HIV/AIDS in common. It will help them to improve HIV/AIDS surveillance system. The social media data is having a rapid growth now a days. The important social media platforms such as Facebook, Twitter, Instagram etc., are playing a very crucial role in spreading an information quickly. There are so many research works carried on social media data, among which twitter analytics is common. Tweet classification, tweet clustering, topic modelling, spam identification and user's network analysis are some of the research work on twitter data. Tweet classification is to categorize the tweets based on the pre-existing class. Tweet pre-processing, feature extraction and feature decomposition are the steps carried out before tweet classification. In order to get a better classification result, all the above-mentioned steps have to be carried out efficiently. In this work an efficient feature extraction method is proposed which increases the outcome of classification accuracy.

2 LITERATURE REVIEW

There are some researchers who done analytics on tweets related to HIV/AIDS. Those works indicate that now a day's social media is vital source for gathering public information.

Rene Clausen Neilsen et al., [13] in their research work analyzed twitter data to inform communication campaigns to promote information related to HIV/AIDS. They also tried to inform about HIV discrimination towards key population to the HIV epidemic. They collected tweets from Brazil and filtered by four categories. Their results encouraged the use of social networking data for improved messaging in campaigns.

- *V Mageshwari is currently pursuing Ph.D in Department of Computer Science, Periyar University, India. E-mail: maheejasmine2290@gmail.com*
- *Dr. I. Laurence Aroquiaraj is Assistant Professor in Department of Computer Science, Periyar University, India. E-mail: Laurence.raj@gmail.com*

Sean D. Young et.al., [14] analyzed Twitter data to identify and track HIV risk behaviors. They tried to map the origin in which the HIV risk behaviors occurred. They used machine learning algorithms such as Logistic regression, Random Forest and Ridge Regression Classifier to classify the tweets based the categories. They used 10-fold cross validation method to examine the speed and accuracy of these models in applying that knowledge to detect HIV content in social media.

3 METHODOLOGY

The proposed work consists of five major steps such as tweet collection, pre-processing, normalization, feature extraction and classification.

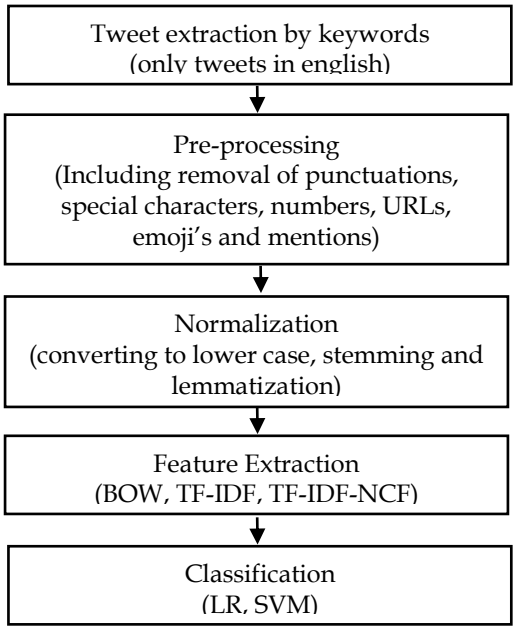

Figure 1. *Research Steps*

4 DATASET DESCRIPTION

Twitter provides an opportunity for the researchers to access twitter data for their work. In order to access twitter data, the user has to create a twitter developer account. The data analytics tools such as R and Python are used to scrape twitter data by mentioning the credentials obtained while creating developer account. Tweets can be collected by mentioning the keywords of interest. In this proposed work,

one lakh tweets are extracted by the keyword HIV/AIDS. Python is the tool used to carry out the proposed work.

5 PRE-PROCESSING

One of the major challenges in Twitter mining is pre-processing [12]. Since twitter data is unstructured it is very vital task to clean the data before moving to further steps. The list of preprocessing steps are

- Eliminating the unwanted special characters and URLs [11].
- Eradication of numbers and punctuation marks
- Altering the misspelled words with manually annotated corpus
- Avoiding redundancy by removing re-sharing and re-tweets
- Stop words are not that much important for classification task. Even it does not provide any important meaning. So, eliminating stop words could help in reducing the dimensionality of the corpus [16].

6 NORMALIZATION

After pre-processing the essential step is to normalize the tweets in order to get the desired text for further feature extraction and classification task. The below are some of the normalization steps carried out in this work,

- Texts are converted to lower case. Since the lowercase and uppercase letters both provide the same meaning, we can convert all the text to lowercase which could be easy for further processing [11]
- Stemming and Lemmatization – stemming of a word would result in removal of prefixes or suffixes with a word. This results in dropping a word to its root forms. Lemmatization is same like stemming, except one difference that it ensures that the word is dropped to its root word belongs to the language. Lemmatization reduces the words to its original dictionary form [12].
- Tokenization – it is the approach used to separate the sentence into small chunks such called tokens. The tokens are fed into the next step as an input. These tokens are just chunks of words [2].

7 FEATURE EXTRACTION

7.1 BOW MODEL

The BOW (Bag of Words) model finds how many times the word is occurred in a document. It is as simple as that it finds the frequency of a word in that document. For example, consider that there are three documents,

D1: My name is Rahul and my student name is Anand
D2: My daughter's name is Lini
D3: My son's name is Barath

After pre-processing step, the BOW model creates a table like matching where each word is matched with its corresponding document. Each document is a column and each word is a row, and each cell is the frequency of word.

	D1	D2	D3
Anand	1	0	0
Barath	0	0	1
Daughter	0	1	0
Rahul	1	0	0
Lini	0	1	0
My	2	1	1
Name	2	1	1
Student	1	0	0

Table 1. Output of BOW model

This method is a very basic feature extraction method in field of text mining. The BOW model just converts the text into vectors which can be represented easily. But the drawback of this model is it does not give any weightage to text. Instead it just calculates the occurrence of the words in each document.

7.2 TF-IDF

Tf-idf means Term frequency-inverse document frequency. It is a weighting scheme frequently used in text mining [3][4]. It is statistical measure which helps to assign a weighting scheme to a word. The tf-idf scheme works better than the basic BOW model, by finding how significant a word is to a document by giving weightage to words in the document.

The tf-idf method usually calculate a weight to each word which in turn helps to find the priority of the word in the corpus [4]. The first step in tf-idf is same as BOW model, where each word is converted to a numerical value.

$$(tf-idf)_{ij} = tf_{ij} * \log\left(\frac{D}{d_j}\right) \quad (1)$$

Where, tf_{ij} denotes the frequency of word j in document i, D represent the total number of documents and d_j denotes the number of documents in which the term i appears. There is a small downside in this formula. If D becomes equal to d_j then $(tf\text{-}idf)_{ij}$ will get equaled to zero. In order to avoid that we can add some smoothing techniques which will help to improve the formula as below,

$$(tf-idf)_{ij} = \log(tf_{ij} + 1.0) * \log\left(\frac{D+1.0}{d_j}\right) \quad (2)$$

The above function will help in smoothing the tf-idf weighting function. This will ensure that the result will not become zero.

7.3 Enhancing TF-IDF

A new in-class parameter has been introduced to overcome the drawbacks of TF-IDF. Usually TF-IDF will calculate weightage of a word to the whole document only. To improve this, we have introduced a new parameter which calculates the term weightage for the document and also the term weightage inside the class. We renamed this as TF-IDF-NCF, where NCF means new class function. The formula of this new weighting method is based on the equation (2) and it is written as,

$$(tf-idf)_{ij} = \log(tf_{ij} + 1.0) * \log\left(\frac{D+1.0}{d_j}\right) * \frac{d_{cij}}{D_{ci}} \quad (3)$$

Where d_{cij} represents the sum of documents in which the word j appears inside the same class c for which the document i belongs to. The next parameter D_{ci} represents the sum of the documents inside the same class c for which the document i

belong to.

8 FEATURE SELECTION

CHI square statistics is a feature selection method, which measures the correlation between class and feature [7]. Let N be the total number of the training samples, A be the times both class c and feature t exist, B be the times where class c doesn't exist and feature t exists and D be the times both feature t and class c doesn't exist. Then CHI square statistics can be written as,

$$x^2(t,c) = \frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (4)$$

9 CLASSIFICATION

Classification is the task of assigning the given data into pre-defined classes. In this work 1,00,000 tweets have been taken in which 70% (70,000) is taken for training and 30% (30,000) for testing. The training sample is assigned five categories such as,

Class 1 – tweets about prevention and awareness
Class 2 – tweets about symptoms
Class 3 – test and care
Class 4 – medicine and treatment
Class 5 – others

In training the model tweets in which the user tweeted about prevention, awareness and about some precautions are labeled as class1. In class2 the tweets about symptoms are classified, in class3 the tweets related to test and care are classified, in class 4 the tweets about medicines, side effects and treatment are classified and in class5 the tweets other than the mentioned above are classified.

9.1 Logistic Regression

Now a day's Logistic regression is one of the commonly used supervised algorithm for text classification [9]. It is used for both binary and multi-classification purpose. If we generalize the logistic regression algorithm for classifying the data into more than two classes, then it is called as multinomial logistic regression. In multinomial logistic regression the Softmax Function is used.

$$softmax(z) = e_z \sum ki = 1e_{zi} \quad (5)$$

9.2 SVM

SVM is a classifier appropriately defined by a separating hyperplane. The algorithm gives output an optimal hyperplane which is capable of classifying test data. SVM classification best suits for classifying text data. It is because in spite of the sparsity of text data, SVM has a tendency to correlate with each other and commonly organized into linearly separable categories[15].

10 EVALUATION MEASURES

The evaluation measures such as precision and recall are calculated. Precision refers to the percentage of the result which are relevant. Recall refer to percentage of total relevant results correctly classified by algorithm.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$Accuracy = \frac{TP+TN}{Total} \quad (8)$$

Table 2. Results of Classification

Feature Extraction Methods	Logistic Regression	SVM
BOW	64.6%	58.1%
TF-IDF	68.2%	65.5%
TF-IDF-NCF	71.9%	70.3%

11 CONCLUSIONS

Social media mining has become a very important research topic now a days. Classification of tweet is one of the significant topics among them. In this work all the necessary pre-processing and normalization methods are carried out to clean the tweets very efficiently. It helped in reducing the dimensionality of the tweets. This work employs a TF-IDF-NCF technique which is enhanced feature extraction method for classifying tweets. Two classification algorithms such as Logistic Regression and SVM are implemented to examine the performance of feature extraction techniques. The results show that the proposed feature extraction method yields higher classification accuracy when compared with the existing ones.

REFERENCES

[1] Akrini Krouska & Christos Troussas, "The Effect of Preprocessing Techniques on Twitter Sentiment analysis", Research Gate, July 2016, DOI:10.1109/IISA.2016.7785373.
[2] Amit G. Shirbhate, Sachin N. Deshmukh, "Feature Extraction for Sentiment Classification on Twitter Data", International Journal of Science and Research, ISSN: 2319-7064, Volume 5 Issue 2, February 2016.
[3] Ammar Ismael Kadhim, Yu-N Cheah, "Improving TF-IDF with Singular Value Decomposition (SVD) for Feature Extraction on Twitter", 3rd International Engineering Conference on Development in Civil & Computer Engineering Applications, 2017, ISSN: 24096997
[4] Ankita Pal, "Principal Component Analysis of TF-IDF In Click Through Rate Prediction", International Journal of New Technology and Research (IJNTR), ISSN: 2454-4116, Volume-4, Issue-12, December 2018, pp 24-26.
[5] Arjun Srinivas Nayak & Ananthu P Kanive, "Survey on Pre-Processing Techniques for Text Mining", International Journal of Engineering And Computer Science, ISSN: 2319-7242, Volume 5 Issue 6 June 2016,Page No. 16875-16879.
[6] Bholane Savita & Prof.Deipali Gore, "Sentiment Analysis on Twitter Data Using Support Vector Machine", IJCST, Volume 4, Issue 3, May-Jun 2016.
[7] Aymen Abu-Errub, "Arabic Text Classification Algorithm Using TF-IDF and Chi Square

Measurements", International Journal of Computer Applications, ISSN: 0975-8887, Volume 93 – No 6, May 2014.

[8] Emma Haddi & Xiaohui, "The Role of Text Pre-processing in Sentiment Analysis", Information Technology and Quantitative Management (ITQM2013), Procedia Computer Science 17 (2013)26-32.

[9] Indra S.T, "Using Logistic Regression Method to Classify Tweets into the Selected Topics", ICACSIS, IEEE, 2016.

[10] Mageshwari V, Dr I. Laurence Aroquiaraj, "Big Data in Health Care Revolution – A Survey", International Research Journal of Engineering and Technology, Volume 3 Issue 9, September 2016, ISSN 2395-0056.

[11] V. Mageshwari, Dr.I. Laurence Aroquiaraj, "Social Media Mining for Analyzing HIV/AIDS – A Preliminary Study", IJIACS, ISSN: 2347-8616, Volume 6, Issue 9, September 2017.

[12] V Mageshwari, Dr.I. Laurence Aroquiaraj, "The Importance of Text Pre-Processing in Twitter Mining", International Journal of Scientific Research in Computer Science Applications and Management Studies, ISSN: 2319-1953, Volume 7, Issue 4, July 2018.

[13] Rene Clausen Nielsen, "Social Media Monitoring of Discrimination and HIV Testing in Brazil, 2014-2015", AIDS Behav (2017) 21:S114-S120, DOI: 10.1007/s10461-017-1753-2

[14] Sean D. Young, Wenchao Yu, "Towards Automating HIV Identification: Machine Learning for Rapid Identification of HIV-related Social Media Data", J Acquir Immune Defic Syndr, February 01 2017, 74(Suppl): S128-S131, doi: 10.1097/QAI.0000000000001240.

[15] Yassine AL AMRANI, Mohammed LAZAAR, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis", The First International Conference on Intelligent Computing in Data Sciences, Procedia Computer Science 127 (2018) 511-520.

[16] Tajinder Singh and Madhu Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis", IMCIP-2016, Procedia Computer Science 89 (2016) 569-554.

[17] ZHANG Yun-tao, "An improved TF-IDF approach for Text Classification", Journal of Zhejiang University SCIENCE, ISSN: 1009-3095, 2005.