Feature Selection Methods for Mining Social Media

V Mageshwari, I. Laurence Aroquiaraj

Abstract: ***People can share their thoughts and opinion through Social Media which can easily widespread. So many public issues and political views are also discussed on social media. HIV/AIDS is also one of the important topics discussed. This work aims to classify HIV/AIDS related twitter data. Since the twitter data is highly dimensional, it is essential to do reduce dimensionality of the data to attain better classification results. Tweets are collected using keyword search and necessary pre-processing steps are carried out. Then feature extraction methods such as Bag of Words (BOW) model and TF-IDF are implemented. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) techniques are used for dimensionality reduction. Finally, classification is carried out and the results are discussed.***

Keywords: Tweets, Pre-processing, BOW, TF-IDF, SVD, PCA,Classification

I. INTRODUCTION

AIDS (Acquired immunodeficiency syndrome) is a syndrome caused by a virus called HIV (Human immunodeficiency virus). Since this syndrome is associated with stigma, most of the people do not discuss about this in-person. But due to the advancement of technology, people are able to share their opinion about this syndrome on social media. The communication campaigns are making effort to reduce the spread of AIDS. They are actively engaged in creating awareness among people. The HIV/AIDS information shared on social media is also a vital source for the campaigns. This information will benefit the communication campaigns to attain their motive of eradicating HIV/AIDS by 2030 [18].

The data on social media is growing tremendously. Instagram, Twitter and Facebook are some of the remarkable social media platforms. A lot of public issues are discussed on these platforms. There are so many researchers who make an effort to mine and gather information from social media. There are even few research works carried on analysing HIV/AIDS information shared on social media [17].

Twitter is a social media platform where the user can share short message of length 280 characters. The big challenge for the computer society is to mine and extract knowledge from this unstructured data. Twitter clustering, classification and topic mining are some of the important research topics. Tweet classification is to classify the tweets based on predefined class labels [16].

In this work AIDS/HIV related tweets are classified. The necessary pre-processing, feature extraction and feature selection method are carried out and better accuracy on classification results are obtained.

II. LITERATURE REVIEW

Some of the research works are carried on analysing HIV/AIDS related tweets. Those works shows the promising role of social media in providing information regarding healthcare.

Rene Clausen Neilsen et.al.,[19] collected tweets generated in Brazil. They examined HIV/AIDS related tweets to promote knowledge about what people speak about this syndrome. This work helped to provide information to communication campaign. They also analysed HIV/AIDS discrimination tweets. This work supported the utilization of social media data for improved messaging in campaigns.

Sean D. Young et.al., [20] in their research work mined Twitter data to identify and track risk behaviours of HIV. They tracked the location in which the HIV risk behaviours tweets occurred. The risk behaviour tweets are classified using algorithms such as Random Forest, Ridge Regression Classifier and Logistic regression. 10-fold cross validation technique is used to examine the accuracy of the work. This work helped to detect HIV related content in social media.

III. DATA DESCRIPTION

Twitter is a social media platform which allows its users to post and share real-time messages called tweets [16]. Now a days we could find a remarkable research works carried out using twitter data. There are few datasets available in online for sentiment analysis of tweets. But if we are taking a particular topic or theme then we can use Twitter API to collect tweets. By using the Twitter API, we can mention the language of the tweet and also, we can extract the related tweets by mentioning the keywords which we need for our research work. Since Twitter data is unstructured it involves little difficulties in carrying analysis on it.

For our research work we have collected 1,00,000 tweets by using the Twitter API. The keyworks we mentioned is ‘HIV’ and ‘AIDS’. The language we preferred is English. The tweets extracted should be pre-processed in order to give better accuracy results. Python and RStudio are the popular tools for twitter mining. In this work python is used to extract tweets and do pre-processing, feature selection and classification on HIV/AIDS tweets.

Revised Manuscript Received on November 10, 2019.

***** Correspondence Author

V Mageshwari*, Department of Computer Science, Periyar University, Salem, India. Email: maheejsamine2290@gmail.com

Dr. I. Laurence Aroquiaraj, Department of Computer Science, Periyar University, Salem, India. Email: laurence.raj@gmail.com

IV. METHODOLOGY

There are six major steps involved in the proposed work as shown in the figure below.

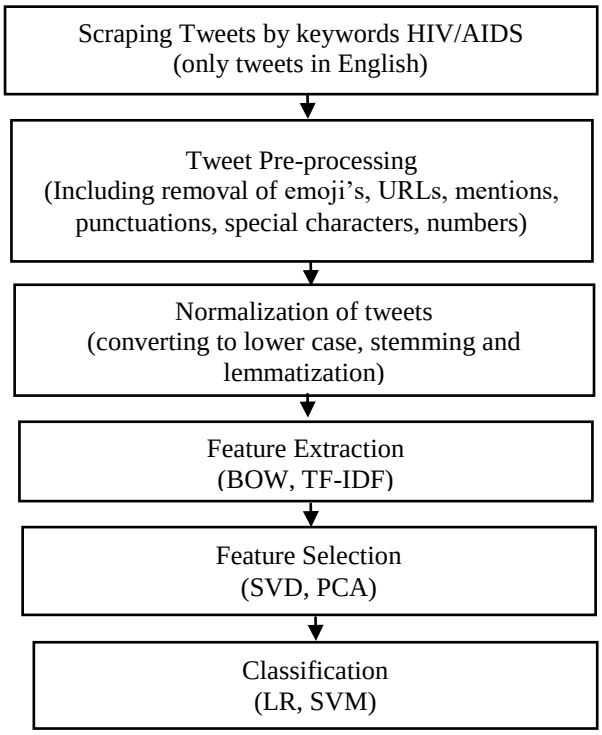

Fig. 1. Methodology Diagram

V. PRE-PROCESSING STEPS

Pre-processing is one of the crucial tasks in twitter analytics. The twitter data is unstructured so pre-processing the tweets is very important to get better accuracy in further steps. The important pre-processing steps carried out in the proposed work includes,

- Getting rid of numeric: Since we are going to process text data, the numeric in tweets are not going to contribute anything to the analytics
- Removal of punctuation marks
- Eliminating URLs
- Eradicating unnecessary special symbols
- Modifying the words which are misspelled
- Removal of Stop words: Since stop words does not add any information to twitter analytics, it is better to eliminate the stop words. By doing so we can reduce the dimensionality of the corpus [22].

VI. NORMALIZATION OF TWEETS

Normalization of tweets has been done to get the desired text which can be appropriate for further analysis. The normalization steps which are carried out in this work is given below.

- Since the uppercase and lowercase words mean the same thing, converting the tweets to lowercase will make ease the further steps.
- Stemming – stemming process is used to trim either the prefix or suffix of the word. As the result the word will be dropped to its root word [17].
- Lemmatization – this process is similar as stemming, but the only change is that the word is reduced to its original dictionary form [18].
- Tokenization – this approach is used to divide the sentence into small fragments called as tokens. These tokens are given as input to the further processing [18].

VII. FEATURE EXTRACTION

The feature extraction step is carried out after the converting the text to tokens. The basic text feature extraction method Bag of Words Model. The second method used in this proposed work is TF-IDF method.

A. Bow Model

The Bag of Words model is a very basic feature extraction method which is used to find the frequency of the word in a document. For example, let us assume that there are three document d1, d2 and d3 as below,

D1: AIDS is not a disease
D2: HIV is a virus
D3: AIDS is caused by HIV virus

If we apply BOW model for the above three documents then the resultant document will be,

D = {AIDS, is, not, a, disease, HIV, virus, caused, by}

Table 1. Output of BOW Model

	D1	D2	D3
AIDS	1	0	1
Is	1	1	1
Not	1	0	0
A	1	1	0
Disease	1	0	0
HIV	0	1	1
Virus	0	1	1
Caused	0	0	1
By	0	0	1

The BOW model changes the text in the sentence into individual vectors. However, there is a disadvantage in this method that it does not provide any weightage to the word, instead it just accounts the occurrence of the words in each document.

B. TF-IDF

Term frequency-inverse document frequency is a weighting scheme used in text mining [7].

It is a statistical method which is used to give weightage to each word in a document. This model is a statistical model. This model performs better when compared with the above model, because it provides an advantage of assigning a weight to the terms in a document [5]. This method also helps in calculation the priority of the word in a document. The initial process of TF-IDF is same as BOW model in which each term is modified to a vector value [23].

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

Where,
j = word,
i = document,
D = complete set of documents,
tf_{ij} = term j frequency in a document i,
d_j = the set of documents in where word i appeared
There is a slight problem with this formula, that if D and d_j become equal then the output answer will be zero. So, some smoothing technique can be added to improve the formula,

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

By modifying the formula in such a way can ensure that the answer will be non-zero.

VIII. FEATURE SELECTION

Feature selection is a method for choosing a subset of appropriate features from the total feature set. This method is carried out to select only relevant features which are more reliable for classification task.

A. Singular Value Decomposition

The singular value decomposition (SVD)is a matrix factorization method [2]. This method acts as an important part in many linear algebra algorithms. The SVD method is used for complex as well as real values. The SVD matrix equation is,

$$A = USV' \quad (3)$$

the above equation tells that if A be a rectangular matrix which can further be decomposed into other matrix components:
U has the orthonormal eigenvectors which belong to AA', U'U = I, V has orthonormal eigenvectors which belong to A'A, S is taken as a dimensional diagonal matrix. This diagonal matrix has the square root of eigenvalues of V or U.
The SVD offers a strong mathematical foundation for the field of text analytics [3]. In SVD, the matrix A represents the text document as a high dimension vector space model. The matrix A is considered as a word x document matrix. The role of SVD in text mining is that it takes high dimensional data and converts it to low dimensional data. By doing so, it can better find the original structure of the data. This SVD method can also be used to reduce redundancy and noise in the data by creating new data dimension from the old one. In text analytics, SVD does the below clarification:

- In U, the words are denoted as rows
- In V, the documents are denoted as rows
- By probing the VS, the similarity of the document can be determined
- By probing the US, the similarity of the word can be determined.

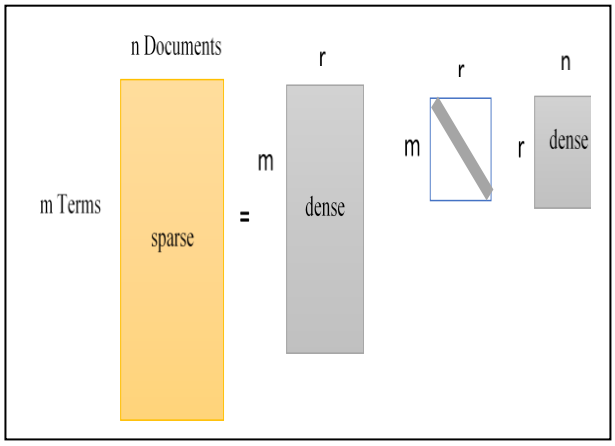

Fig. 2. Reduced SVD for Term Document Frequency Matrix

B. Principal Component Analysis

PCA is used to reduce dimension of the data from high-dimensional dataset to low-dimensional space [26]. It is considered as a linear data projection method [25]. In this work PCA is used to map the features extracted from TF-IDF method. PCA tries to preserve variance in the low-dimension space. PCA gets initiated with computing the following covariance matrix.

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)})\,(x^{(i)})^T \quad (4)$$

IX. STEPS INVOLVED FEATURE EXTRACTION AND SELECTION PROCESS

The below steps are carried out in sequence involving both the feature extraction and selection process.
Step 1: Categorize each tweet with a keyword set
Step 2: Transform all the tweets into vectors which can be represented in numeric form, then label the keyword by category ID
Step 3: Estimate the existing TF-IDF method and smoothened TF-IDF method
Step 4: Carry out feature extraction process on the numeric data
Step 5: Compute the accuracy based on the confusion matrix.
Step 6: Carry out the process same as Step 1 to Step 4. Then do the feature selection process using the dimensionality reduction technique SVD and PCA
Step 7: Compute the confusion matrix accuracy for the process involved in step 6.
Step 8: Compare the accuracies achieved in both the step 5 and step 7.
Step 9: Repeat the above process by updating the keyword set.

X. CLASSIFICATION

In classification task the input data is partitioned into pre-defined classes [18]. For this proposed work, 1,00,000 tweets are scraped by using Twitter API.

The extracted tweets are divided into training and testing set. We have taken 70% for training and 30% for testing. The training data is labelled with five classes,

C1 – the tweets related to prevention and awareness are labelled as C1 category

C2 – the tweets in which people spoke about symptoms and other related issues are categorized in C2

C3 – the tweets in which people mentioned about testing & care are considered in C3

C4 – the tweets about treatment and medicine are categorized under C4

C5 – the tweets which does not fall under the above-mentioned classes are categorized under this class C5.

After labelling the tweets with concerned classes, the classification algorithms such as SVM and Logistic Regression are carried out to test the model.

A. Logistic Regression

The Logistic regression can be used for both the regression and classification task. In classification it can be applied to perform both binary and multiple classification. For using the logistic regression for classifying data into multiclass then a Softmax Function is used [18]. Consider that there are k classes, then every class will have its own parameter θ_k, which maps x to $\theta_k^t\, x$ as given in below equation.

$$p(y = k|x) = \frac{\exp(\theta_k^T x)}{\sum_{i=1}^{K} \exp(\theta_k^T x)} \qquad (5)$$

B. Support Vector Machine

The SVM classification can be applied to both binary and multiple classification [20]. For multi-classification two SVM methods can be used,

OVA: One versus All method is used to fit the k class as, $\hat{f}_k(x), k = 1, \ldots K$. Which classifies x^* to the class for which the equation is largest.

OVO: One versus One method will be used to fit all $\binom{k}{2}$ pairwise classifiers, $\hat{f}_{kl}(x)$ which classifies x^* to the class for which the class wins most pairwise competitions.

XI. EVALUATION MEASURES

Confusion matrix has been used as an evaluation metric in this work. The overall accuracy is calculated by using precision and recall as shown below.

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

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

$$Recall = \frac{TP}{TP+FN} \qquad (8)$$

Table II: Accuracy of Classification without Feature Selection

Feature Extraction Method	Logistic Regression (in %)	SVM (in %)
TF-IDF	68.2	65.5
Log TF-IDF	70.8	69.9

Table-III: Accuracy of Classification with Feature Selection

Feature Extraction	Feature Selection	Logistic Regression (in %)	SVM (in %)
TF-IDF	SVD	72.5	66.2
Log TF-IDF	SVD	73.1	70.1
TF-IDF	PCA	74.8	72.7
Log TF-IDF	PCA	75.3	73.9

XII. RESULT AND DISCUSSION

The results from the above section shows that the algorithm can do better classification by using the SVD method for feature selection. Since text data is highly dimensional, it is always better to perform some dimensionality technique before performing classification task. The comparison chart is plotted for the accuracy results obtained.

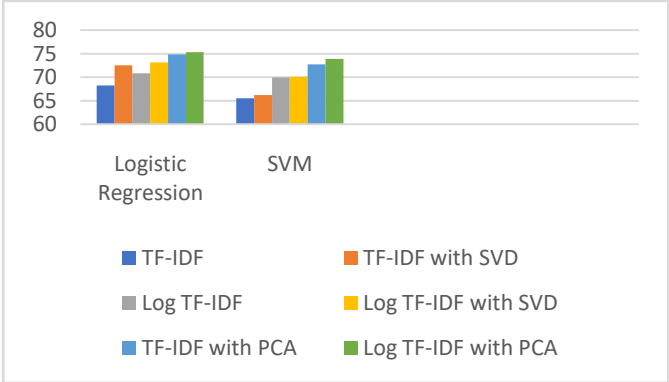

Fig.3. Comparison among the classification task with and without Feature Selection

The chart shows that with a feature selection technique, the classification algorithm could yield a better accuracy. The SVD and PCA methods are used as dimensionality reduction technique in many research works. In the field of text mining it is used for dimensionality reduction and also for feature selection process which obviously results in improved classification accuracy.

XIII. CONCLUSION

Classification of social media data has become a trend now a days. In earlier, usually people make use of social media data for market basket analysis. This is because we utilized social media platform mostly for sharing our reviews. But now a days this has been changed. People started sharing their opinion, ideas and thoughts. We could also notice that health related issues are also discussed on social media. Mining these kinds of health-related social media data can help the communication campaign to improve the surveillance system. This work shows the necessary steps in classifying the HIV/AIDS related tweets. Further dimensionality reduction techniques are also implemented to show how it

can help in enhancing the classification task.

XIV. FUTURE WORK

The following could be considered as a future work,

- The n-gram construction can be done
- More feature selection algorithms can be implemented
- The accuracy can be measured and compared using different classification algorithms

REFERENCE

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 Gupta, "Sentiment Analysis of Tweets using Machine Learning Approach", International Journal of Computer Science and Mobile Computing, Vol.6 Issue 4, April-2017.
5. 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.
6. 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.
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. Bholane Savita & Prof.Deipali Gore, "Sentiment Analysis on Twitter Data Using Support Vector Machine", *IJCST*, Volume 4, Issue 3, May-Jun 2016.
9. Carl A. Latkin, "Social network approaches to recruit, HIV prevention, medical care and medication adherence" J Acquir Immune Defic Syndr. 2013 June.
10. Dr. S. Vijayarani, "Preprocessing Techniques for Text Mining -An Overview" International Journal of Computer Science & Communication Networks, Vol 5(1), 7-16.
11. 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.
12. Fouzi Harrag, "Improving Arabic Text Categorization Using Neural Network with SVD" Journal of Digital Information Management, Volume 8 Number 4, August 2010.
13. Hyunsoo Kim, "Dimension Reduction in Text Classification with Support Vector Machines", Journal of Machine Learning Research 6(2005) 37-53.
14. Indra S.T, "Using Logistic Regression Method to Classify Tweets into the Selected Topics", *ICACSIS*, IEEE, 2016.
15. 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.
16. 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.
17. 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.
18. V. Mageshwari, Dr. I. Laurence Aroquiaraj, "An Efficient Feature Extraction Method For Mining Social Media", International Journal of Scientific & Technology Research, October 2019.
19. 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
20. 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.
21. 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.
22. Tajinder Singh and Madhu Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis", *IMCIP-2016*, Procedia Computer Science 89 (2016) 569-554.
23. ZHANG Yun-tao, "An Improved TF-IDF approach for Text Classification", *Journal of Zhejiang University SCIENCE*, ISSN: 1009-3095, 2005.
24. K Subba Reddy, "Using Reduced Set of Features to Detect Spam in Twitter Data with Decision Tree and KNN Classifier Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019.
25. Kristine Mae M. Adlaon, "Dimensionality Reduction of Feature Word Vectors for Sentiment Classification of Philippine Political Related Tweets" Proceedings of the 18th Phillipine Computing Science Congress (PCSC 2018)
26. Jian-lin LI, "A Text Classification Algorithm Based on PCA" 2017 2nd International Conference on Computer Science and Technology (CST 2017), ISBN: 978-1-60595-461-5

AUTHORS PROFILE

V. Mageshwari has completed M.SC and M.Phil in Department of Computer Science and Engineering from Bharathiar University, Coimbatore, Tamil Nadu, India in 2015. She is pursuing Ph.D Degree in Department of Computer Science, Periyar University, Salem, Tamil Nadu, India. She has attended a total of nine workshops and seminars. She has attended more than six conference in which presented paper in four among them. She published nine articles in International journals. Her area of interest includes Big Data, Image processing, Data Mining, Natural Language Processing and Text Mining. She was graduated as a Big Data Hadoop Architect from Simplilearn in the year 2017. She received Elite certificate with silver medal by NPTEL in 2019.

Dr. I. Laurence Aroquiaraj is working as an Assistant Professor in The Department of Computer Science, Periyar University, Salem. His Educational Qualifications includes M.Sc., M.Tech., M.Phil., M.C.A., and Ph.D in Computer Science. He has fourteen years of Teaching and Research experience. He also has a three year of Industrial experience. His main area of Research is Medical Image Processing. He is also much interested in the areas including Data Mining, Internet Technologies and Networking. He attended fourteen International Conferences. A total of ten Conferences, Seminars and Workshops are organized by him. Overall forty-three research papers are published by him in International Journals and Conference Proceedings. Total twenty-three National Journals and Conference Proceedings are published by him. He is a member of Professional Bodies such as CSI (Life Member), IEEE, IAENG, IET Image Processing, Elsevier-ACI, IRG-IJCT and IJCII. He has extended his guidance to hundred and four M.C.A. and twenty M.Sc. (Computer Science) students. He also guided thirty-four M.Phil., (Computer Science) scholars. Eight Ph.D. (Computer Science) scholars are pursuing their research work under his guidance. He immensely worked on UGC funded project entitled, "Early Detection of Breast Cancer from Mammogram Using Soft Computing Techniques".