Journal of Engineering Sciences

Vol 11, Issue 5,May/2020
ISSN NO:0377-9254

STATISTICAL ANALYSIS OF ONLINE DATA

Dr.Ch. Surya Kiran[1], K. Anusha[2], K. Trinath reddy[3], Sk. Afshan[4]
1 Professor, NRI Institute of Technology, 2, 3, 4 Scholars, NRI Institute of Technology

ABSTRACT: Event detection and classification has become an important task in social media platforms. It facilitates the exploration and navigation of events with early preventive action plans. The main challenges are the features of short / conversational, heterogeneous and live social media data. Online social network apps like Facebook, Weibo, have played a key role in the lives of people. Tweets contain tremendous data. But it's a difficult problem how to mine the tweets and get valuable information. This paper focuses on collecting numerous tweets from twitter, preprocessing the tweets and the tweet classification [1] [2] into specific categories by using Decision tree classifier.
Keywords: Event detection, Social media, Classification.

1. INTRODUCTION

Detection and classification of events in social media is different from detection and classification in other text format. The rationale why detection of events in social media streams is tougher is brief and noisy content, diverse and rapidly changing topics, and enormous volumes of knowledge. Many people within the world use social media to remain connected to their friends, relations and colleagues through their computers and mobile phones. Due to their real-time existence, social media like Facebook, Snapchat, Whatsapp and Twitter have recently received tons of attention. The proliferation of social media exposes many opportunities for research. For several reasons, social media information might be used, like monitoring accidents, predicting events, and even early warning systems. Messages posted on Twitter revealed everything from everyday tales to the new news and events at the local and global level. Event detection and classification has become a crucial task in social media platforms. It facilitates the exploration and navigation of events with early preventive action plans. The most challenges are the features of short / conversational, heterogeneous and live social media data. Online social network apps like Facebook, Weibo, have played a key role within the lives of individuals. Now days there's rapid development of social media platforms and Twitter is one among the social media's most famous. Twitter may be a micro-blogging platform where users express views that are called tweets within the sort of 140 characters and short messages. Many tweets are being created a day. Such enormous data has introduced new areas for educational and business-related research work. Identification of Twitter messages and identification of twitter emotions are the 2 main areas during which many researchers work round the world. Tweets contain tremendous data. But it is a difficult problem the way to mine the tweets and obtain valuable information. This paper focuses on collecting numerous tweets from twitter, preprocessing the tweets and therefore the tweets are classified into specific categories by using Decision tree classifier.

2. RELATED WORK

As Twitter is one of the most famous platforms for people to express their ideas and opinions, it has given a lot of scope to conduct research and find out the topics on which people are discussing a lot. There are a lot of ways and techniques to classify the tweets and find the topics on which people are interested in.

Some of the current popular classifiers are support vector machine (SVM), neural network (NN), KNN, Naïve Bayes, j48 and so on, are built in an inductive learning way. Some of the classification techniques that are being used to classify the twitter messages are Support Vector Machine (SVM) and Naïve Bayes classifiers [3] [4].

The existing systems even though they are serving the purpose, there are some disadvantages of these approaches

- The execution time of SVM is more and efficiency and accuracy are less when compared with Decision tree classifier.
- The structure of SVM algorithm is very complex and does not work properly if the

data set is very big.

- Naïve Bayes does not give proper accuracy if the data set is small.
- There is a chance of loss of accuracy in the case of Naïve Bayes classifier.

3. METHODOLOGY

Methodology of this proposed model is of two phases namely preprocessing, and classification.

Preprocessing is the process of preparing the data for algorithm. Since the data for our project is collected from the Twitter platform, it should be preprocessed before it is given to the Decision Tree Classifier.

- First, the URL's, links, HTTP tags should be removed.
- Tweets from twitter contain lot of incomplete words and a lot of unnecessary words. They must be removed if unnecessary or completed before they are given to the Decision Tree Algorithm.
- Even the tweets contain lot of stop words like ('is', 'are', 'to' … etc) which are of no use when it comes to classifying the tweets into their respective domains. So they must be removed before processing.

For this Preprocessing, the concept of Natural Language Processing (NLP) [5] is used. The module that is used is nltk.
(natural language tool kit) in Python. After preprocessing the data, the data must be splitted.

In splitting phase, the data splits into training set and testing set. Training set is used to train the model and testing set is used to check the accuracy or error rate of the model.

Classification phase involves constructing a classification model (Decision Tree Classifier here) and predicting the classes to which the tweets belong to using that model.
For classification purpose, Decision Tree Classifier is used.

Decision tree is a popular and powerful tool

Used for classification purpose. It is a flow chart like structure. It classifies all the instances by sorting them from the root to some leaf node which provides the classification of the instance.

- Internal nodes denote a test on attribute,
- Branch nodes denote outcome of the test,
- Leaf nodes denote or represent the class label.

Construction of Decision Tree

A tree can be constructed by splitting the source set into subsets. Those subsets are splitted based on an attribute value test. This process is repeated in a recursive manner on every derived manner. The process is referred to "recursive partitioning".
The recursion is completed when splitting no longer adds value to the predictions or the subset at a node all has the same value of the target variable.

Attribute Selection Measure: An attribute selection measure is a used for selecting the splitting criterion that separates a given data partition in a "best" way, D, of class-labeled training tuples into individual classes.

The attribute selection measure that is used in our project is "Information Gain".

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

- pi: nonzero probability that an arbitrary tuple in D belongs to class Ci and is estimated by |Ci, D|/|D|.
- A log function to the base 2 is used; the reason is that information is encoded in bits.

For individual attribute, information gain is calculated as

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

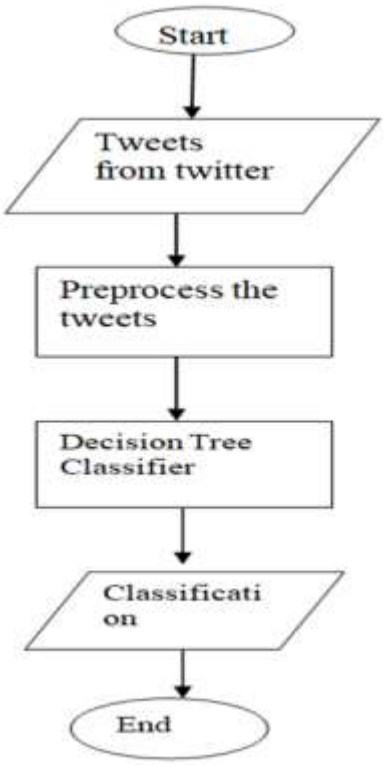
- |Dj | / |D| acts as the weight of the jth partition.
- InfoA (D): Expected information required to classify a tuple from D based on the partitioning by A.

Overall information gain is calculated as

$$Gain(A) = Info(D) - Info_A(D).$$

Attributes with highest value of Information gain is considered as splitting criterion.

FLOW DIAGRAM



4. RESULTS

In our project, the dataset used is twitter dataset which consists of tweets related to Nepal earth quake.

```
['disaster management' 'disaster management' 'disaster management'
 'disaster management' 'disaster management' 'disaster management'
 'assistance' 'disaster management' 'disaster management'
 'disaster management' 'bilateral help' 'disaster management' 'relief'
 'bilateral help' 'distress aid' 'disaster management'
 'disaster management' 'disaster management' 'bilateral help'
 'disaster management' 'disaster management' 'disaster management'
 'distress aid' 'assistance' 'disaster management' 'disaster management'
 'disaster management' 'disaster management' 'relief'
 'disaster management' 'assistance' 'assistance' 'bilateral help'
 'disaster management' 'disaster management' 'disaster management'
 'disaster management' 'disaster management' 'assistance' 'relief'
 'disaster management' 'relief' 'assistance' 'disaster management'
 'disaster management' 'disaster management' 'disaster management'
 'disaster management' 'friendly relations' 'bilateral help'
 'disaster management' 'disaster management' 'disaster management'
 'assistance' 'relief' 'assistance' 'disaster management'
 'disaster management' 'rescue' 'rescue']
```

```
the accuracy with Decision Tree Classifier is:
0.4
disaster management    17
relief                  6
distress aid            5
assistance              2
```

Fig: Classification of tweets using Decision Tree

```
disaster management    20
distress aid            2
bilateral help          2
rescue                  2
relief                  2
support                 1
friendly relations      1
dtype: int64
the accuracy with Naive Bayes Classifier is:
0.03333333333333333
```

Fig: Classification of tweets using Naïve Bayes

5. CONCLUSION

In this project, Decision Tree Classifier has been proposed. Decision tree has improved performance over other classifiers that are now being used. In general, decision tree classifiers exhibits good accuracy. Decision tree induction algorithms are used for classification in many application areas like medicine, manufacturing and production, financial analysis.

6. FUTURE ENHANCEMENT

The conducted experiments showed an honest performance of Decision Tree Classifier and achieved an honest overall accuracy. In future accuracy of an equivalent are often improved with the assistance of neural networks.

REFERENCES

[1]. Twitter Trending Topic Classification Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary

Department of Electrical Engineering and Computer Science Northwestern University, Evanston, IL 60208 USA Email: {kml649, drp925, ran310, mpatwary, ankitag, choudhar}@eecs.northwestern.edu

[2]. W. Zhang and F. Gao, Procedia Engineering An Improvement to Naive Bayes for Text Classification,‖ vol. 15, pp.
2160–2164, 2011.

[3]. Short Survey on Naive Bayes Algorithm
Pouria Kaviani, Mrs. Sunita Dhotre
M.Tech student, Department of Computer Engineering, Bharati Vidyapeeth University, College of Engineering, Pune
Associate Professor, Department of Computer Engineering, Bharati Vidyapeeth University, College of Engineering, Pune.

[4]. P. Selvaperumal and A. Suruliandi, "A short message classification algorithm for tweet classification," 2014 International Conference on Recent Trends in Information Technology, Chennai, 2014, pp. 1-3.

[5]. Natural Language Processing: State of The Art, Current Trends and Challenges
Desha Khurana, Aditya Koli, Kiran Khatter and Sukhdev Singh
Department of Computer Science and Engineering
Manav Rachna International University, Faridabad-121004, India
Accendere Knowledge Management Services Pvt. Ltd., India