INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 3)
Available online at: www.ijariit.com

Classification of tweets into various categories using classification methods

Shubham
shubm0939@gmail.com
Dayananda Sagar College of Engineering, Bangalore, Karnataka

Shashank Kumar
shashanksingh8877@gmail.com
Dayananda Sagar College of Engineering, Bangalore, Karnataka

Sunanda Dixit
sunanda.bms@gmail.com
Dayananda Sagar College of Engineering, Bangalore, Karnataka

Piyush Kumar
piyushdbg2013@gmail.com
Dayananda Sagar College of Engineering, Bangalore, Karnataka

ABSTRACT

Social media has become an important part of our regular life and Twitter is one of the famous among them. As the growth and uses of social media are increasing rapidly so does the twitter, the number of Twitter users has reached an estimate of 330 million monthly active users. Twitter provides a list of trending topics in real time, but it is often hard to understand what these trending topics are all about. It is important and necessary to classify these topics into various categories with high accuracy for better information retrieval. With the enormous volume of data being generated on TWITTER [1], it is imperative to find a computational means of filtering. To address this problem, we classify tweets into various categories such as sports, politics, technology, etc. We will use various algorithms such as Naïve Bayes classifier, Support Vector Machine classifiers to classify the tweets into various categories and check the accuracy of each algorithm.

Keywords: *Naïve Bayes, SVM [2], Twitter, Classification.*

1. INTRODUCTION

Twitter is a micro-blogging site where users express their views related to various fields such as sports, entertainment, politics etc. Message generated on Twitter are called tweets, which is of at most 140 characters long. The number of Twitter users has reached an estimate of 330 million monthly active users. Not only regular users but also celebrities, company representatives, politicians and even country presidents are audiences of twitter. Every day millions of tweets are generated containing a huge amount of information. These tweets are very important in understanding what the trending topics are. Twitter message classification [3] is one of the important areas of research related to tweets. It is important and necessary to classify these topics into various categories with high accuracy for better information retrieval. Classification of tweets and assigning categories to tweets has many applications like spam filtering, sentiment analysis etc.

2. RELATED WORK

There are many existing systems which has been proposed in the recent times. MapReduce [8] is a programming model and an associated implementation for processing and generating large data sets. A map function is specified by the user that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. The model is easy to use, even for programmers without experience with parallel and distributed systems, since it hides the details of parallelization, fault-tolerance, locality optimization, and load balancing. A large variety of problems are easily expressible as MapReduce computations. MapReduce is suitable only for batch processing jobs. It does not do well for graph, iterative, incremental and many other kinds. Applications that involve pre-computation on the dataset brings down the advantages of MapReduce. Another system is of **Subjective genres** [10], such as "editorial", are often one of the possible categories. Other work explicitly attempts to find features indicating that subjective language

is being used. Techniques for genre categorization and subjectivity detection can help us recognize documents that express an opinion. By checking the message or tweet by seeing it, we can classify it into one of the several categories such as sports or law etc. It is not useful as well as accurate for a large number of texts as well as for those messages which have contents related to various categories as it is not accurate.

3. APPROACH

To automatically classify tweets from Twitter of various types based on predefined categories. We took various categories into consideration for classifying twitter data. These categories are business, motorcycles, space, medicines, religion, politics, sports and technology. The data was collected from various sources as shown below:

- **Input Data**: The real-time data consisting of the user tweets.
- **Training dataset**: Fetched from 20 News Group sets.
- **Final Deliverable**: It will return a list of all categories to which the input tweet belongs.

There are some concepts used in this proposed method to get the desired result. These are:

- **Outliers removal**: It is used to remove low frequent and high frequent words using Bag of words approach.
- **Stop words removal**: It is used to remove most common words such as "the", "is", "at", "which", and "on".
- **Keyword Stemming**: It is used to reduce inflected words to their stem, base or root form using porter stemming.
- **Cleaning crawl data**: The crawl data is being cleaned.
- **Spelling Correction**: It is used to correct spellings using EDIT DISTANCE method.
- **Named Entity Recognition**: It is used for ranking result category and finding the most appropriate result.
- **Synonym form**: It is used when a feature of test query is not found as one of
  dimension in feature space then replaces that word with its synonym.

The proposed method uses two algorithms for classifying tweets into various categories. These three algorithms are Naïve Bayes classifier[5], rule-based classifier and support vector machine which are discussed in following parts.

A. Naïve Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector x = (x1,x2,…..,xn) representing some *n* features it assigns to this instance probabilities

p(Ck | x1,x2,…..,xn) for each of *K* possible outcomes or classes.

Using Bayes' theorem, the conditional probability can be decomposed as

p(Ck | x) = p(Ck) p(x | Ck) / p(x). In plain English, using Bayesian probability terminology, the above equation can be written

Posterior = prior * likelihood / evidence.

The features of the words are selected and then this technique is applied to extract features to classify them into a particular category. Tokenization is carried out to split a stream of text into smaller units called words or phrases. The steps of the algorithm are discussed below:

- **Data Streaming**: The collected raw tweets are applied as an input to produce the particular category.
- **Pre-Processing of Extracted Data**: It cleans the unstructured textual data into a structured textual class by removing the punctuations and additional symbols [6].
- **Feature Selection**: It extracts features of a tweet.
- **Naïve Bayes Classifier**: It is used to predict the probability of a given word to belong to a particular class.

Pre-processed data is given symbols as input to train input set using Naïve Bayes [7] classifier and that trained model is applied to the test to generate a particular category.

B. Support Vector Machine

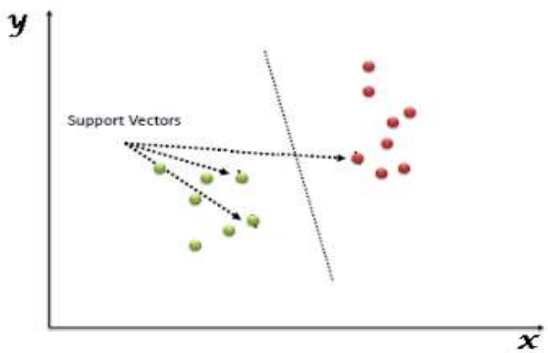
The numeric input variables (x) in your data (the columns) form an n-dimensional space. For example, if you had two input variables, this would form a two-dimensional space.

A hyper plane is a line that splits the input variable space. In SVM, a hyper plane is selected to best separate the points in the input variable space by their class, either class 0 or class 1. In two-dimensions, you can visualize this as a line and let's assume that all of our input points can be completely separated by this line. For example:

B0 + (B1 * X1) + (B2 * X2) = 0

Where the coefficients (B1 and B2) that determine the slope of the line and the intercept (B0) are found by the learning algorithm, and X1 and X2 are the two input variables.

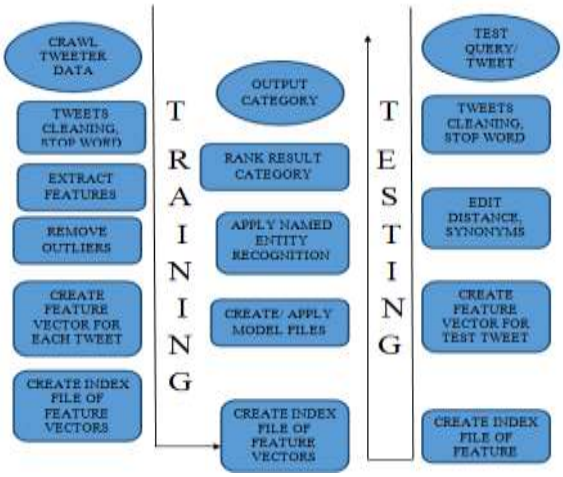
We can make classifications using this line.



- Above the line, the equation returns a value greater than 0 and the point belongs to the first class (class 0). 2.
- Below the line, the equation returns a value less than 0 and the point belongs to the second class (class 1).
- A value close to the line returns a value close to zero and the point may be difficult to classify.
- If the magnitude of the value is large, the model may have more confidence in the prediction.

4. ARCHITECTURE

A system architecture is a conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. A system architecture can comprise system components that will work together to implement the overall system. We have used the bag-of-word technique for data representation. The bag-of-words [4] model is a simplifying representation used in natural language processing. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [9]. The below figure shows a general block diagram describing the activities performed by this project.



We propose a new model for classification of tweets into various categories using scikit learn and conda. We have used Bags of Words approach for extracting features from the words. Two of the classification algorithms are used, namely Naïve Bayes, and SVM to classify the tweets, and showing the accuracy of each of the method. We have created a pipeline for easy implementation of the code, and then have used it.

5. CONCLUSION

We came to a conclusion that SVM is more accurate than the Naïve Bayes classifier. Since Twitter has stopped giving data for any use, we have to use data manually by feeding it to the classifier to classify it. Feeding the data manually, we can't get a system which can classify any data given to it. In future, if Twitter starts giving data, then a system can be made which will help to train the classifiers with any data, and we will be able to classify them.

Also, with certain improvement in the classifiers, we can attain more accurate results. This will be very helpful in the future, as there is more spam coming nowadays, so to avoid those spam, we need a filter, and this system can be transformed as one. We used two different algorithms for classifying tweets into various categories.

6. REFERENCES

[1] P.Selvaperumal and Dr.A.Suruliandi "A Short Message Classification Algorithm for Tweet Classification", 2014 International Conference on Recent Trends in Information Technology.

[2] InoshikaDilrukshi, Kasun De Zoysa, Amitha Caldera," Twitter News Classification Using SVM, the 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013

[3] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in IEEE ICDM Workshops (ICDMW). IEEE, 2011.

[4] A.Tripathy, A.Agrawal, S.K Rath, "Classification of sentiment reviews using n-gram machine learning approach,"Journal of Expert Systems With Applications(Springer), Vol. 57, pp. 117- 126, 2016.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in ACM SIGIR. ACM, 2010.

[6] Bo Pang and Lillian Lee, "Thumbs up? Sentiment Classification using Machine Learning Techniques".2002.

[7] B.liu, E.Blash, Y.chen, G.chen, D.shen, "Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier", Journal of International Conference on Big Data (IEEE), Oct. 2013.

[8] Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters" In a 6th symposium on operating system design and implementation, Vol- 6, pp.137-149, 2004.

[9] H. Takemura and K. Tajima, "Tweet classification based on their lifetime duration," in Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012, pp. 2367–2370.

[10] E.Berger, 2009, "This sentence easily would fit on Twitter: Emergency physicians are learning to Tweet", Annals of Emergency Medicine, volume 54, number 2, pp. 23A-25A.

