IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 5, No. 1, March 2016, pp. 41~44
ISSN: 2252-8938

Twitter Tweet Classifier

Ashwin V
Departement of Information Technology, SRM University

Article Info

Article history:

Received Dec 4, 2015
Revised Feb 7, 2016
Accepted Feb 25, 2016

Keyword:

Microblogging
Tweets
Stemming
Machine learning algorithm

ABSTRACT

This paper addresses the task of building a classifier that would categorise tweets in Twitter. Microblogging nowadays has become a tool of communication for Internet users. They share opinion on different aspects of life. As the popularity of the microblogging sites increases the closer we get to the era of Information Explosion.Twitter is the second most used microblogging site which handles more than 500 million tweets tweeted everyday which translates to mind boggling 5,700 tweets per second. Despite the humongous usage of twitter there isn't any specific classifier for these tweets that are tweeted on this site. This research attempts to segregate tweets and classify them to categories like Sports, News, Entertainment, Technology, Music, TV, Meme, etc. Naïve Bayes, a machine learning algorithm is used for building a classifier which classifies the tweets when trained with the twitter corpus. With this kind of classifier the user may simply skim the tweets without going through the tedious work of skimming the newsfeed.

Corresponding Author:

Ashwin V,
Dept. of Information Technology,
SRM University,
Kattankulathur Chennai, India.
Email: 94.ashwinvee@gmail.com

1. INTRODUCTION

Successful microblogging services such as twitter have come to an integral part of daily life of millions of Internet users. Interest in mining in twitter increased with the widespread use of these services. Because they become the ware house of peoples opinion on current issues. Twitter is the most popular and the second most used social networking site. Since its launch in 2006, the popularity of its use has been drastically increasing, that more than 100 million user post 340 million tweets a day in 2012. As of march 2016, Twitter has more than 310 million monthly active user.

In twitter, users are allowed to create status messages which are called tweets. Tweets are of 140 characters which are written by registered twitter users about their life, opinions on variety of topics and discussions on current issues. As more and more users share their opinion on several fields, their view on current issues, microblogging sites become the assets of user's opinion and sentiments.

In this research we use a dataset formed by collecting twitter tweets. These tweets are collected and are fed to the classifier which classifies them into several categories like Sports, News, Entertainment, Politics, and Meme. With the help of these categories user may just have to choose the category of his/her interests.

2. RELATED WORK

With the increase in popularity of social networks and blogs, analysis and mining has become a field of interest for many researchers. There are a lot of researches done in twitter such as user classification,

Trend detection, Sentiment classification, trend detection etc. But this research mainly carries out the task of segregating tweets into several categories for user's convenience.

In overviewing L.Huag, R.Bhayani and A.Govind, 2009 research, they addressed the classification of tweets based on sentiment. These tweets are classified into either positive or negative with respect to query term. They have also used the emoticons occurring in tweets as "silver" labels i.e., labels with more uncertain status than the ones found in usual "gold" standards for tweet sentiment analysis. This research mainly benefits consumers who want to know about the sentiment of products before purchase, or companies that want to monitor the public sentiment of their brand. While J.Read in his research (Read, 2005) he has used emoticons such as ":-)" and ":-(" which serve as noisy labels and formed a training set for sentiment classification. For this purpose, the author collected texts containing emoticons and divided the data set into positive and negative samples

Sankaranarayanan et al. 2009, built a news processing system for identifying the tweets that corresponds to late breaking news. They collected the tweets and they removed noises from them. They clustered the tweets using the clustering algorithm called leader-follower clustering, which allows for clustering in both content and time. The other issue which is addressed in this research is identifying relevant locations associated with the tweets. Pang and Lee, 2002 researched the performance of various machine learning techniques like Naive Bayes, Maximum entropy and SVM in the specific domain of movie reviews. By this they were able to achieve an accuracy of 82%.

Sriram et al. 2010, their work is more relevant to ours. They classified tweets into predefined set of classes such as news, events, opinion, deals and private messages with the use of information about the author and also feature which are extracted from tweets such as "@username", shortening of words, slang, etc. They classified tweets in order to improve information filtering. Their feature outperformed the bag. Of word model approach in the classification of tweets.

Different from the above works, this research defines a system that would classify all the tweets irrespective of the trend, sentiment, user into categories like News, Meme, Sport, Entertainment which would make twitter far more convenient to use and also saves ample amount of time for the user.

3. SYSTEM DESIGN

Our system here is segregated into different modules.

3.1. Tweet Retrieval Module

The basic requirement for the retrieval of Tweets is the Twitter API. Registration for the API is done using an existing Twitter account. Once registered, the user is provided with a Consumer Key, a Consumer Secret Key, an Access Token and an Access Token Secret using which the tweets are retrieved from the user's timeline.

3.2. Text Processing Module

To analyse and classify text, there are certain pre-requisite actions that must be performed. The retrieved tweets are written onto a Text File. Each tweet must be written on to a different text file, all of which must be in the same directory. These text files are the Documents that will be used. The documents are first put through the process of cleaning. Cleaning refers to the removal of any and all punctuation marks from within the document. Once the documents are cleaned, the next process is to remove the Stop Words. Stop Words are words like articles, pronouns, prepositions and conjunctions which would not affect the eventual classification. Then, stemming which is the process of extracting the root word from each of the words in the document is carried out. Stemming returns a set of root words that can then be fed into the classifier.

3.3. Conversion Module

The conversion that has to occur is to get an ARFF File to be fed into Weka i.e., Waikato Environment for Knowledge Analysis which is a Machine Learning tool developed at the University of Waikato in New Zealand. The whole directory of Documents is converted into one ARFF File using the Weka's Core.Converters Library. The conversion is done using the *TextDirectoryLoader* method that derives from the *TextToArff* class in Weka. The command *weka.core.converters.TextDirectoryLoader dir <Directory Path> > <ARFF File Path>* must be entered in Weka's Command Line Interface which carries out the conversion.

3.4 Classifier Module

We build a classifier using the Naïve Bayes classifier. Naıve Bayes classifier is based on Bayes' theorem.

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)} \quad (1)$$

Where d is the Document, and c is the Category.
The best class in NB classification is the most likely or *maximum a posteriori* (MAP) C_{MAP} class:

$$\text{CMAP} = \text{argmax} \in \text{C P(d1, d2, ... , dn} \mid \text{c)P(c)} \quad (2)$$

Document d is represented as feature d1, d2…, dn.

The classifier works after the Attributes (Classification Feature) are mentioned on the ARFF File. Of the documents in the whole set, 70% are used as the training set, i.e., these 70% of the tweets are manually labelled as Sports, News, Entertainment, Technology, Music, TV, Meme, etc. and the remaining 30% is used as the Test set.

3.5. User Module

This can be called the front-end of the application. This includes the Graphic User Interface (GUI). This provides the link between the application and the user. This the module where the user will be displayed with tweets that are categorised as Sports, Entertainment, News, Meme, Private Message.

4. RESULT

This graph shows the accuracies per classes using Naïve Bayes classification. This graph summarizes the result of our classifier. The accuracies in classification into catgories are pretty high which is very good. The classification of tweets into categories are only based on the words contained in the tweets that we used in the training set. These words may contradict in more than one categories which will affect the accuracy. But in our case that impact has been very low except for the category Others because the ambiguity in tweets that fall under the others category is very high and so they fall into some other category resulting in this low accuracy. There are several approaches like 8F, BOW, 9F that can be used to increase the accuracy.

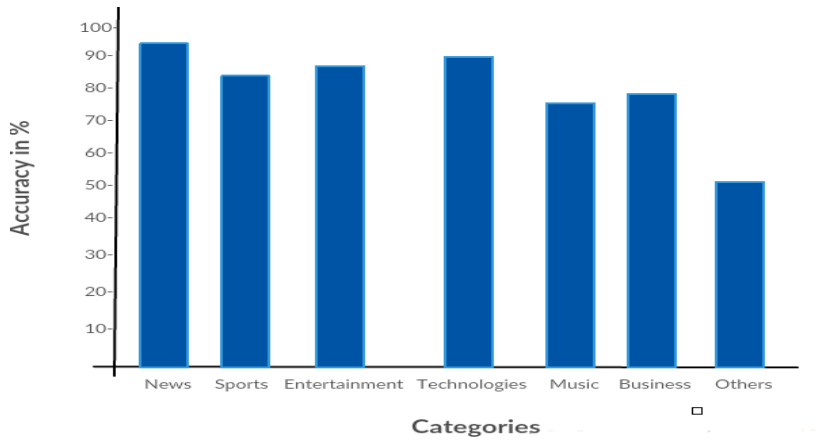

Figure 1. Accuracies per category using Naïve Bayes

5. FUTURE ENHANCEMENT

Mission learning techniques perform well for classification of tweets. We believe that accuracy could still be improved. More advanced comparison approaches can be taken into consideration such as

clustering. And also using machine learning algorithms like SVM, Maximum entropy can be used in addition to the Naïve Bayes classifier to improve the efficiency in classification.

As mentioned earlier, twitter has 500 million tweets every day, hence in near future this mindboggling number is bound to increase and subsequent categories need to be accordingly implemented.

6. CONCLUSION

As discussed earlier microblogging nowadays has become one of the major modes of communication among internet users. A recent research has identified it as online word-of-mouth branding (Jansen et al., 2009). The large amount of information contained in microblogging web-sites is what makes them an attractive source of data for mining and analysis. Although twitter messages have unique characteristic compared to other corpora, machine learning algorithm have shown to classify tweets with similar performance.

In this paper we use twitter feeds as our data set and categories them based on their nature and significance of tweet. Further researched is needed to continue to improve the accuracy in difficult domain. Machine learning algorithm can achieve high accuracy for classifying when using this method.

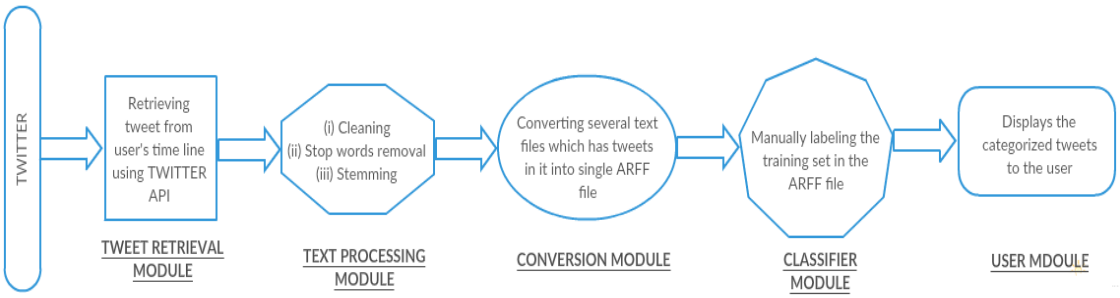

Figure 2. System Design

REFERENCES

[1] J Sankaranarayanan, H Samet, BE Teitler, MD Lieberman, J Sperling. Twitterstand: news in tweets. 2009.
[2] B Sriram, D Fuhry, E Demir, H Ferhatosmanoglu, M Demirbas. Short text classification in twitter to improve information filtering. 2010.
[3] A Go, R Bhayani, L Huang. Twitter sentiment classification using distant supervision. 2009.
[4] M Bush, I Lee, T Wu. NLP-based approach to Twitter User Classification. 2010.
[5] J Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. 2005.
[6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter Trending Topic Classification. 2009.
[7] J Benhardus. Streaming Trend Detection in Twitter. 2010.
[8] B Pang, L Lee, S Vaithyanathan. Sentiment classification using machine learning techniques. 2002.
[9] S kabir and M Tahmim. Localized twitter opinion mining using sentiment analysis. 2009.
[10] Ted Pedersen. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. 2000.
[11] YS Yegin Genc, JV Nickerson. *Discovering context: Classifying tweets through a semantic transform based on Wikipedia*. Proceedings of HCI International. 2011.
[12] TM Mitchell. Machine Learning. McGraw-Hill, New York. 1997.
[13] Weka 3: Data Mining Software in Java, http://www.cs. waikato.ac.nz/ml/weka/.
[14] Naïve Bayes classifier. Retrieved from https://web.stanford.edu/class/cs124/lec/naivebayes.pdf
[15] Naive Bayes text classification. Retrieved from http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html.

BIBLIOGRAPHY OF AUTHOR

Ashwin V is from Chennai born on 05/10/1994 Has completed Undergraduate B.Tech degree in Information Technology at SRM University Chennai, India in the year of 2016. And currently working at Emerio Technologies, Guindy, Chennai, India.
This is the Author's first research work.