Volume 4, Issue 4, April 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Sentimental Analysis of Theme Trending on Social Network

Rajesh Yadav[1], P. Shanthi Bala
Department of Computer Science
Pondicherry University, India

Abstract: This paper is about the sentiment analysis of theme trending on social network such as twitter which analyses the given text into genres like politics, sports, entertainment, technology, health and business. It also reviews the user's interest. In this paper, we trained set of predefined tweets using supervised machine learning approach. It makes decision regarding the genres which it belongs to. In this paper, the tweet messages are categorized according to the latest user interests and in which genres it belongs to. It is achieved using a supervised machine-learning approach based on support vector machine (SVM) and Naive Bayes approach. The accuracy of the classifier after constructing the data set is more than 80%.

Keyword: Sentimental Analysis, Social Network, Twitter, Tweet Classification, SVM

I. INTRODUCTION

Sentiment analysis or opinion mining is used to identify and extract subjective information with the use of natural language processing, text analysis and computational linguistics. The sentimental analysis is accomplished using various social networks. Twitter is an online social networking and micro-blogging service that enable users to send and read "tweet" in which text messages are limited to 140 characters. Registered users can read and post tweets but unregistered users can only read them. Twitter was created in March 2006 by Jack Dorsey and by July 2006, the site was launched. The service rapidly gained worldwide popularity with 500 million registered users in 2012 who posted 340 million tweets per day. The service also handled 1.6 billion search queries per day. It has been described as "the SMS of the internet". Nowadays, social networks like twitter are the latest trend in the globalized world.

Twitter is used in different scenarios by a broad set of different users. A person who access to receive the updates is called follower on twitter, blogs, and other social networking sites. Twitter profile page has "Following aaa", where aaa is the number of people that person is following and "Followers aaa", where aaa is the number of people that are following that person. Twitter is used by many famous persons all over the world. The U.K Prime Minister Gordon Brown, the UN Secretary General Ban Ki-moon, the president of the European Commission Jose Manuel Barso, and the U.S president Barack Obama, with 2.6 Million followers. Actually twitter provides API to access its data.

Currently, many research works is going on twitter for sentiment analysis. For example, Huberman and Sitaram[1] used twitter to guess the box office revenues for upcoming movies and achieved 97% accuracy. In U.S. presidential debate in 2008[2], twitter is used to predict the political election results thus Obama won over McCain. Jansen et al. [3] showed that 19% of micro blogs mention a brand among which 20% contain sentiments. Using sentiment detection, a market analyser can analyse his/ her products in market competition, whether the product is good or bad. The sentiment has been analysed by using a supervised machine learning classifier such as Naive and SVM which can automatically classify the tweets in different category. It is not based on rule or other external knowledge but it will be examine tweets itself using a supervised machine learning method. There is a need of dataset to build a classifier though which it can be evaluated. Our goal is to create a high quality dataset which can be used to classify the tweets with very high accuracy. Due to the restriction of 140 characters in twitter, user can't write more information. So, hash tag is used to give more information to the users. Usually, hash tags bind more information within single tag. When the users required more information it could be searched using the hash tag. For example "I am not happy". #happy means this sentence is tag by #happy. The message is tag with 'happy' and other users are able to search for that tag using the twitter site.

Tweets related to hash tag is shown in Figure 1. Early 90s, the machine learning approach is started and it provides very good results and it also eliminates the need to enter and maintain a rule set. Initially, a training set is prepared using various data sets. Based on these data sets, the process builds an automatic classifier which categorizes the new tweets. It is termed as supervised machine learning approach. In this, the collection of proper distinctive set of tweets is a challenging task. The set of tweets have been utilized to train the model. Earlier, the classes identified are status updates, personal, information sharing, etc but this classification never required any training model as certain attributes of a particular tweet.

The rest of the paper consists of following: Section 2 describes related works. Section 3 describes the proposed architecture. Section 4 describes the algorithm used in supervised machine learning approach. Section 5 is about the experimental setup and results. Finally, the conclusion and future work are presented in section 6.

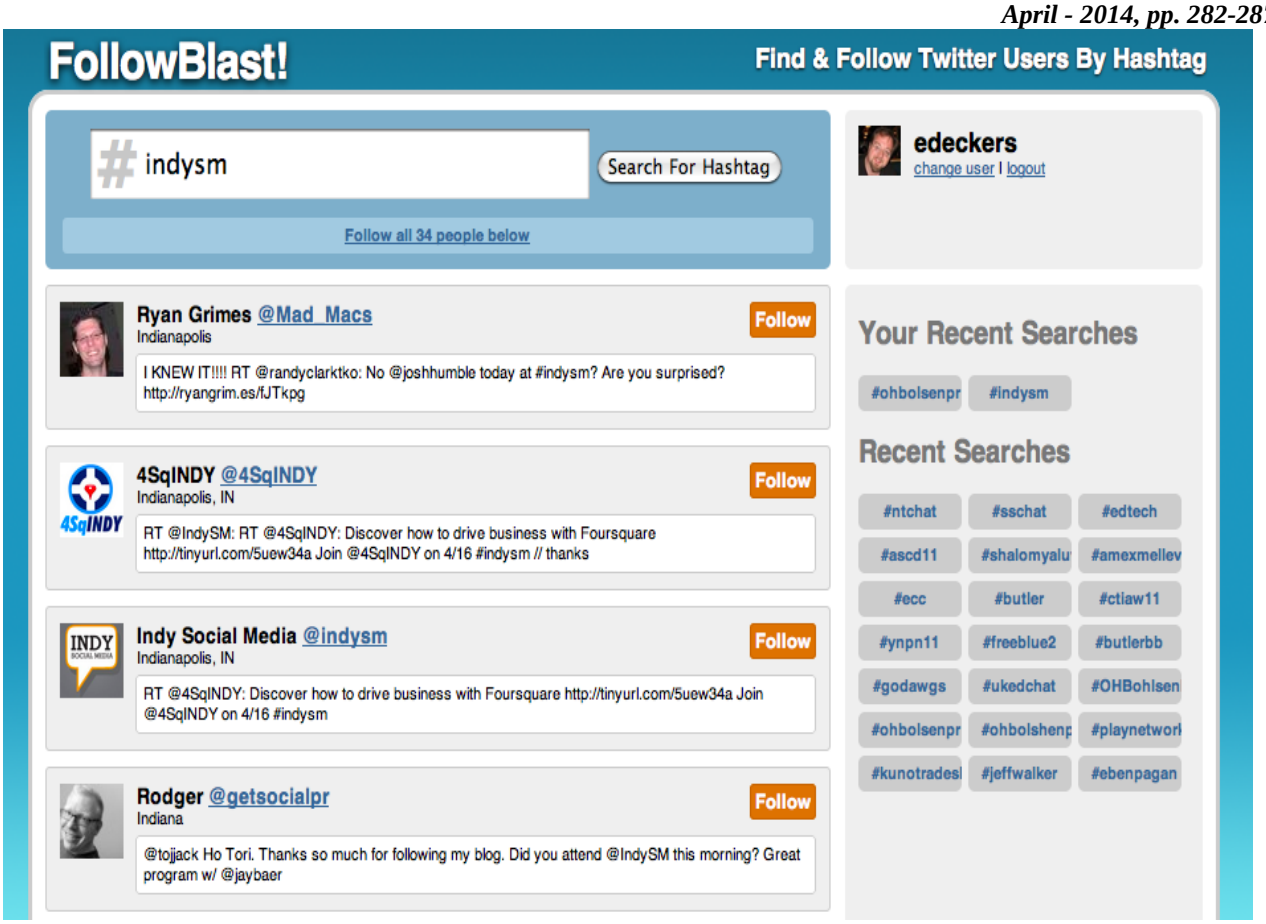

Figure 1. Tweets related to hash tag

II. RELATED WORK

Naaman et al. [4] categorized Twitter messages based on their content. They have developed 9 categories, and manually assigned the latest 10 Tweets of 350 randomly selected users. The categories are Information Sharing (IS), Self Promotion (SP), Opinions/Complaints (OC), Statements and Random Thoughts (RT), It's all about me (ME), Questions to followers QF), Presence maintenance (PM),Anecdote – me (AM), Anecdote – other(AO). The results show that more than 40% of the tweets are categorized as ME, followed by RT, OC and IS with approximately 20% each. In other words, this means that 20% of the Tweets have a news character, while 80% can be characterized as user-to-user communication. Sankaranarayanan et al. [5] proposed a system called `TwitterStand' which captures breaking news of tweets. They have splitted the tweets into two categories (`news' and `junk') and use Naive Bayes classifier to categorize them. Cheong and Lee [6] categorized the twitter users into following groups: 'Personal', 'Group' (e.g. a fan club), 'Aggregator' (e.g. news agencies), 'Satire' and 'Marketing'. Pang et al. [7] evaluated machine learning techniques which is performed on sentiment detection.

Kamps et al. [8] combined SVMs with good semantic differentiation (using WordNet relationships) and lemmatization, and accomplished an accuracy of 89% on the same data set. Kazushi et al. [9]proposed demographic estimation algorithm for profiling twitter users , based on their tweets and community relationships. Jing Guo et al. [10] proposed a flexible stream miming approach for hot topic detection and used frequent pattern stream mining algorithm to detect hot topics from twitter stream. Raymondus Kosala et al.[11] used the twitter information to solve the traffic problems in Jakarta, the capital of Indonesia. Sriram et al. [12] classified the tweets into a predefined set of classes such as events, deal, news, opinions, and messages based on author information and domain specific features extracted from tweets. Genc et al. [13] introduced a classification technique based on Wikipedia. Classification is performed by mapping messages into their most similar Wikipedia pages and calculating semantic distances between messages based on the distances between their closest Wikipedia pages. Kathy Lee et al. [14] classified the Twitter Trending Topics in 18 categories such as technology, sports, health, politics, etc using two approaches the network based classification and Bag of words. They have achieved 65% and 70% accuracy in text based and network based classification respectively. The twitter is used to provide the opinion about the particular product or services based on the feedback given by the users in the twitter.

III. ARCHITECTURE

The architecture of proposed classification system is shown in figure 2. It consists of three modules: Tweet extraction, tweet filtering and training model creation and user interface.The tweet messages are fetched using twitter API and stored the tweets in the database for further processing. The surplus tweets are filtered and an index file is created from the filtered tweets. The extracted tweet messages are trained and categorized using supervised machine learning approach.

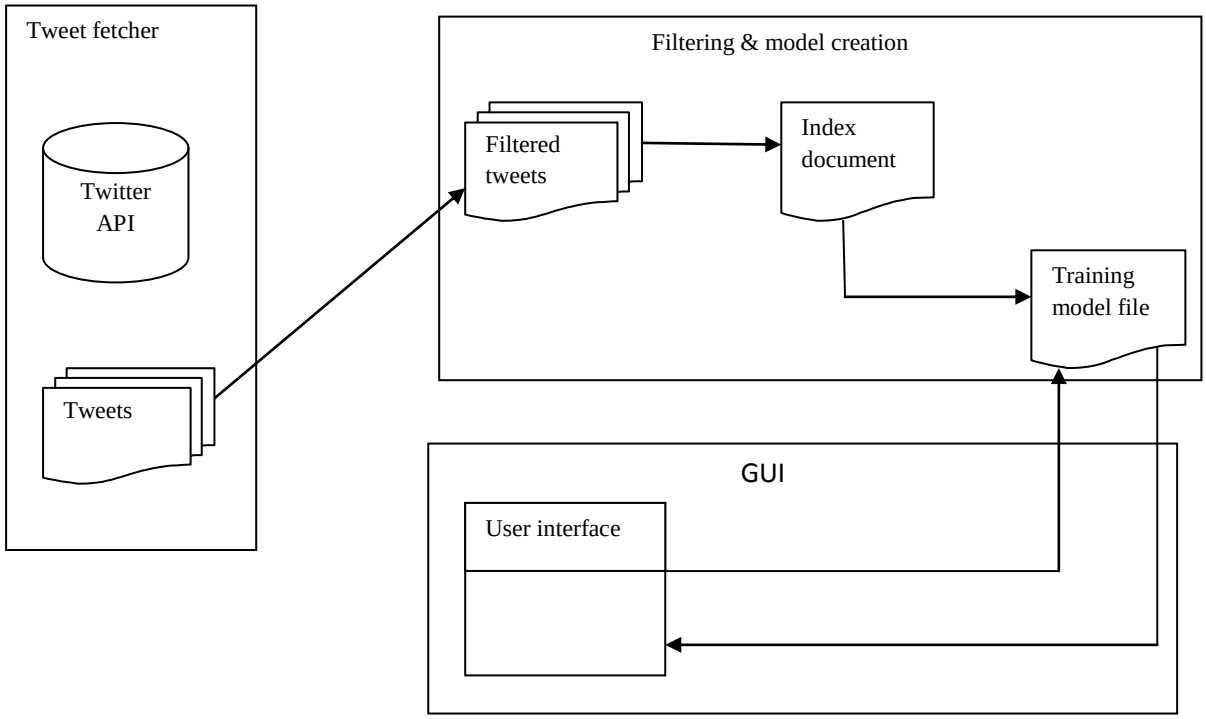

Figure 2. Architecture of tweet classification

A.TWEET EXTRACTION

A set of hash tag is predefined for each class. If a tweet contains any of the hash tag then it falls under that particular class. Tweets are extracted using twitter API. Hash tags for each category are as follows:

1. *Business*: #business, #branding, #BusinessNews, #market, #OilTrader, #budget
2. *Entertainment***:** #oscars, #beyonce, #american idol, #grammys and #BigBangTheory
3. *Sports***:** #olympics, #IPL, #coachfootball, #SkySports, #sports and #yankees
4. *Technology***:** #iphone, #ipad, #skype, #kindle, #google, and #android
5. *Politics*: #10DowningStreet, #Obama2014, #idemocracy, #Political, #ThaiPoliticS, #badpolitics, #GunPolitics and #GunControl

B. TWEET FILTERING & TRAINING MODEL CREATION

In the extracted tweet messages, lot of noise is available. So, it should be cleaned to improve the accuracy, efficiency and scalability of the classification process. The following transformation is done on tweets:

1. *Removal of non-English tweets*
2. *Removal of links*: - Budget from state Senate due out today http://t.co/lt9BxAid5u
3. *Removal of Retweet detail***:** -RT @startupbritain: A third of UK citizens want to start up in #business
4. *Removal of usernames***:-**@FundicaLooking forward to working with you! #business #website #funding
5. *Generalization of amounts***:-**The legislature is debating the $32.7B state budget. -> Replace it with $XX.
6. *Generalization of percentage*:-#News #BusinessNews Monsanto's Profit Rises 22%. -> Replace it with XX%.
7. *Conversion of tweets into lower case*
8. *Removal of stop words*: Removal of common words like a, all, above, and, or, after, you, would, etc.

C. USER INTERFACE

The user can access the functionalities of proposed system using the developed interface.

IV. ALGORITHMS

A. Support Vector Machine

SVM is a technique used in supervised machine learning. Given a set of categories, which contain an arbitrary number of items SVM predict which category a new items belongs to. Due to SVM's Bi-Classification nature, one to one variant of SVM model was created for having multiclass classification. Here nC2 training models are constructed each model votes for its class and the class getting maximum votes gets the selection. The theoretical background of SVMs is explained detailed in [15] and [16].

SVM is used in training and testing the classifier. The following steps have been used for tweets classification using SVM.

1) Training

1. First create Tweeter developer API to get customer key, token, etc to fetch Tweets from Twitter.
2. Used Java API and hash tag i.e. #India, #cricket, #ipl to fetch tweets for different category from twitter like sports, Technology, Entertainment, Politics, Business and Health.

3. Then, apply pre-processing steps
4. Remove all non-English character from tweets.
5. Remove all links from tweets and at same time check if tweets contain more than two links then consider these tweets as spam and remove whole tweets.
6. In next step, create dictionary by collecting unique words from all category.
7. Create index
8. One to all mapping is required to create index i.e. if there are three class A,B,C then map AB, AC, BC, but not map BA because AB is equals to BA.
9. Due to this approach, total number of index file = nC2 where n = number of category for 6 category we have 15 index file.
10. In each index file we label class +1 and -1 i.e. for AB index file use label +1 for all tweets that belong to A and use label -1 for all tweets that belong to B class file.
11. Create Model file for each index file
    a. Use LIB-SVM to create model file.
    b. In this approach, liner kernel is used because we transform multiclass classification to binary class classification, gamma=0 (based on some experiment), svm_type=c_svr… etc

Total number of model file = nC2 where n = number of category.

Testing (for testing we use 10 fold testing)

1. Use 9/10 part for training purpose and 1/10 part for testing purpose
2. In testing 1-3 of training is similar.
3. Create index file
    a. Label all training tweets to some random value from +1,-1.
    b. Create single index file.
4. Give this index file to input to SVM that use previous created all model file to classify each tweets.
5. Based on the output classify tweets into one of these 6 category.
6. Steps 1- 5 are repeated for 10 fold, in which 1/10th part of tweets for testing purpose and remaining 9/10th part for training purpose.

B. NAIVE BAYES

A high dimensional dense vector for each tweet is constructed. Vector is constructed using each unique word of training tweets. Here each word is treated as an independent feature. These features are treated as independent of each other and they contribute equally in classification of any tweet.

In Naive Bayes, each document is represented by a vector $x = (x_1,\ldots X_m)$, where $x_t = 1$ if a certain term t is present in the document, and $x_t = 0$ otherwise (hence the word Binary in the name). Queries are built in the same way. `Independence' indicates that the model assumes that the terms are not associated with each other. Using the Bayes rule, the probability of relevance of a certain document can be calculated. Despite the simplifying assumptions this model uses, it achieves good results in practice [17].

A feature can be defined as an attribute which helps to identify that object. For example, the features of a car are, amongst others, 'vehicle', '4 tires', 'engine', 'moves on ground'. Now, the Naive Bayes classifier assumes that none of those features are related to each other. While this is simplifying assumption, it turned out that this algorithm performs remarkably well in practice, even when strong feature dependencies are present[18]. In Naive Bayes again two phase training and testing. In training phase which is already used in SVM (means all steps will be same) for fetching the tweets. In Naive Bayes the index file is generated and code is written in indexcreatetrain.java, which again call indexbuildtrain.java internally. After that call wekademo.java again for creating model of index file. In this java file there is a function createmodel( ), which will create model.

In Testing phase write our tweets in a file called test, which again goes through indexing and modelling process, for that call indexcreatortest.java and it calls internally indexbuildtest.java.index will be generated. Now it uses the naive. model for classifying the tweets into classifiers, for this call testtweet.java.

V. EXPERIMENTS AND RESULTS

In our experiment, tools such as WEKA and LIBSVM are used. WEKA is most widely used machine learning tool that support different modelling algorithm for data pre-processing, classification, clustering, regression and feature selection. LIBSVM are the popular open source machine learning libraries developed at the National Taiwan University. It supports classification and regression. First setup the data, once the user had been set up, the classifier was evaluated with a K- folds cross validation. For k, a value of 5 was chosen. Given 40 records per category, this resulted in 8 runs. For the first run, the first 5 records of each category where used as test set, and the remaining 35 records as training set. For the second run, the records 5-10 of each category where used as test set, and the records 1-5 and 11-40 as training set. Use cross validation to predict correctly the tweets in each category. For example, a value of 86.80 means that 86.8% of the Tweets in this category have been classified as being a member of this category, hence classified correctly. The other 13.2% have been classified with the wrong category.

The results of the previous section are analysed by a confusion matrix. Confusion matrix is a method which is used in machine learning approach when the number of categories is more than two.
The tweets related to news, user and company shown in table1.

TABLE 1 TWEETS CATEGORY

	News	**User**	**Company**
News	1297	190	109
User	173	1272	94
Company	172	210	1207

For Naive Bayes we got accuracy of 86.66% and for SVM it was 90%. Scaling may improve results but close observation of the classifier behaviour is necessary. The results shown in table 2

TABLE 2 CLASSIFIER BEHAVIOUR

Tweets	**Naive Bayes**	**SVM**
Business	80%	76%
Entertainment	86%	98%
Health	95%	98%
Politics	91.98%	96%
Sports	81.82%	99.6%
Technology	85%	99%

VI. CONCLUSION

In this paper, a supervised machine learning approach is used to classify the tweets in six categories and predict the category of a given tweet with high accuracy of more than 80%. The main challenge is to fetching the tweet and training the machine. Two algorithms are used to classify the tweets and SVM works significantly better than Naive Bayes. In the future work, the training model can be updated using correctly identified tweets.

REFERENCES

[1] S. Asur and B. a. Huberman, "Predicting the Future with Social Media on Web Intelligence and Intelligent Agent Technology,'' Proceeding of the 2010 IEEE international conference. Vol.1, pp. 1492–499, 2010.

[2] N. a. Diakopoulos and D. a. Shamma, "Characterizing debate performance via aggregated twitter sentiment," Proceeding of the 28th International Conference on Human factors Computer System. p. 1195,ACM, 2010.

[3] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," ProceedingJournal of the American society for information science and technology., Vol. 60, pp. 2169–2188, 2009.

[4] M. Naaman, M. Naaman, J. Boase, J. Boase, C. H. Lai, and C. H. Lai, "Is it Really About Me on Message Content in Social Awareness Streams,''Proceedings of the 2010 ACM conference on Computer supported cooperative work. pp 189-192,2010.

[5] J. Sankaranarayanan, B. E. Teitler, H. Samet, M. D. Lieberman, and J. Sperling, "TwitterStand : News in Tweets," Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.vol. 156, pp. 42–51, 2009.

[6] M. Cheong and V. Lee, "Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base,"Proceedings of the 2nd ACM workshop on Social web search and mining. pp.1-8, 2009.

[7] B.Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up ?Sentiment Classification using Machine Learning Techniques,"Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Volume 10. 79–86, pp 79-86, 2002.

[8] J. Kamps, J. Kamps, M. Marx, M. Marx, R. J. Mokken, R. J. Mokken, M. de Rijke, and M. de Rijke, "Words with attitude," Proceeding of the 1st International Conference on Glob, pp. 332–341, 2002.

[9] K. Ikeda, G. Hattori, C. Ono, H. Asoh, and T. Higashino, "Twitter user profiling based on text and community mining for market analysis,''Knowledge-Based Syst., vol. 51, pp. 35–47, 2013.

[10] J. Guo, P. Zhang,JianlongTan, and L. Guo, "Mining Hot Topics from Twitter Streams,"Proceeding of the International Conference on Computational Science. Vol. 9. pp. 2008–2011, 2012.

[11] R. Kosala, E. Adi, and Steven, "Harvesting Real Time Traffic Information from Twitter,"Proceeding of the International Conference on Advances Science and Contemporary Engineering, vol. 50. pp. 1–11, 2012.

[12] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering,"Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval p. 841, 2010.

[13] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting Wikipedia as external knowledge for document clustering," Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, vol. 28, p. 389, 2009.

[14] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter Trending Topic Classification," Preceeding of the11th IEEE International Conference on Data Mining, pp. 251–258, 2011.
[15] V. N. Vapnik, ''The Nature of Statistical Learning Theory,'' Springer, vol. 8. 1995, p. 188.
[16] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition,"Proceeding of the international conference onData Mining Knowledge Discovery, vol. 2, pp. 121–167, 1998.
[17] C. D. Manning and P. Raghavan, "An Introduction to Information Retrieval,"Vol. 1, p. 6, 2009.
[18] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Los," vol. 29, pp. 103–130, 1997.

