Opinion Mining of Real Time Twitter Tweets

Akash Shrivatava
Graphic Era University, Dehradun

Shweta Mayor
Graphic Era University, Dehradun

Bhasker Pant
Graphic Era University, Dehradun

ABSTRACT
Twitter is a real-time information network and micro-blogging service that allows users to post updates. The service rapidly gained worldwide popularity that connects to the latest stories, ideas, opinions, and news. It is a powerful tool for real-time way of communicating with people by combining messages that are quick to write, easy to read, public and accessible anywhere. On Twitter anyone can read, write and share messages or tweets. **Opinion mining** is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Social media plays an important role in inferring the opinion of the authors. In this paper we focused on tweets that will result in analyzing the view of the public on generally discussed topics. A tweets puller is developed that automatically collects random opinions and classifier tool that performs classifications on that corpus collected from Twitter. Our classification is based on features extracted and classified into POSITIVE, NEGATIVE and NEUTRAL. The results further evaluated and concluded to infer the performance of the classification through SVM.

Keywords
Opinion mining, classification, SVM, twitter tweets mining, Data mining, web mining, text categorization, support vector machine.

1. INTRODUCTION
Today the World Wide Web (Web) is a popular and interactive medium for disseminate the information. Web pages designed generally for collection of facts, coined as Content data. Content analysis has ranged between the "Twitter as a serious business" and "Twitter as conversations apparently about nothing" with content analysis tending towards the serious end expressing concerns over the volume of personal and social content [10]. Social networking websites are the epoch of expressing views. Social sites like TWITTER1, FACEBOOK2, FRIENDSTER3 and many more have become a platform for expressing, sharing, articulating one's views, thoughts and expressions. Every discussed topic by the author of those comments, views, thoughts shared is followed by the perception of other users. It may include any political issue, religious issue, technology, product, movie or music review and much more daily gossiping issues flooded in their surroundings [2]. The dataset collected from all these sites can be effectively and efficiently used for marketing, case studies, building relationships and social networking. They can easily point out inferences and draw conclusions about their product, technology or political point whatever they all are concerning with by going through opinions comes from these sites [3]. The need to collect opinions from these social networking sites and draw conclusions that what people like/dislike, has been the most important aspect in today's scenario. The dataset is collected from TWITTER as it contains large number of Tweets or messages concerning users personal thoughts and public views from different regions and countries. TABLE 1 shows typical example of some Tweets. In our paper, we study that how these sites would use for sentiment analysis purposes which not only shown their opinion or point of view towards any matter but also provide their needs, demands from the current scenario. We collected around 5000 tweets from Twitter which is eventually split automatically into three sets as follows: POSITIVE VIEWS: The tweets that can convey the view or thought in the appreciation of some particular subject. NEGATIVE VEIWS: The tweets that can convey the view or thoughts in the disparagement of some particular subject. NEUTRAL VIEWS: The tweets that can convey the views or thoughts in the clinical of some particular subject. We show how to classify these features based on different impact through classifier that extracts features in three separate classes. Finally we use LIBSVM support vector machine tool to train and testing accuracy of system that up to which extent our system does opinion mining.

[1]http://twitter.com

[2]http://facebook.com

[3]http://friendster.com

TABLE1. Example of TWITTER Tweets with user views

pretty great to know we can #makeachange. best part of the job. ThANK YOU to all my fans making a positive difference in the world. love u
#InCollege the BEST news to hear is that -- 1.) class is canceled 2.) paper not due til later 3.) your lowest scores r dropped
RIP to the 96 people who sadly passed away at Hillsborough 23 years ago today. Today isn't about football, it's about remembrance.#JFT96
If anyone is having a bad day, remember that today in 1976 Ronald Wayne sold his 10% stake in Apple for $800. Now it's worth $58,065,210,000
What a day, unbelievable shows today !amazing crowds in sydney. You ozzy's go hard! Love it! #1DdownUnder

1.1 Contribution
The contribution of our paper is as follows.

1. Our methodology helps us to establish the DOMAIN DICTIONARY that contains the feature terms of individual classified files.
2. We have designed Twitter TWEETS PULLER which can pulls 1000 tweets at a time when it is connected to the server site.

3. We develop a CLASSIFIER TOOL that classify collected corpus tweets from twitter into respected category which would automatically store as per their feature in separate files. After the classification computational linguistic analysis is done. We can also build sentiment classification system based on features extraction. We conduct experimental evaluations to produce real time results on a set of real twitter tweets posted to prove that our technique is efficient enough and performs better than previously proposed methods.

1.2 Organizations

The remaining paper is as follows divided into further section. In section 2, we discuss what are the material and tools we have used for the extraction of tweets, training and testing data. In section 3, we give the explanation of approach for collecting the corpora and its classification. Further experimental evaluations performed by LIBSVM shown in section 4. Finally we conclude our paper about our work.

2. RELATED WORK

Opinion mining became a research interest due to extensive usage of social networking web. People tremendously generate data on web in the form of opinion, reviews etc. Opinion mining had been broadly explained in [12]. In this study, authors describe existing techniques developed for retrieval of data from huge repository of web. In opinion mining author of [13] build corpora of web-blog to determine user's mood and to perform sentiment analysis. They had taken emotions sign's as to indicate user sentiment. They used SVM learners for classification of sentiments of users and then analyzed various scenarios to conclude complete sentiment of document. Finally the last statement of document had been considered as the successful strategy. In [14] author used emoticons like smiley's, for happy or sad mood. They collected dataset from UseNet group containing these emoticons for expressing their sentiment. This collected dataset had been classified into two groups 'Positive'and 'Negative' samples. Emoticons trained classifiers i.e. SVM and NAÏVE BAYES were able to obtain upto 70% accuracy test upon this sample datasets.

3. MATERIAL AND TOOL USED

3.1 Data used

In my research work the main focus is on the Twitter tweets. They will be further use for mine opinion on the basis of features contain in the tweets extracted.

SVM implementation- LIBSVM

LIBSVM is a library for Support Vector Machines (SVMs) developed by Vladimir Vapnik [8] at AT&T, SVMs quickly gained attention from the pattern recognition community due to a number of theoretical and computational merits [11]. A typical use of LIBSVM involves two steps: first, training a data set to obtain a model and second, using the model to predict information of a testing data set. LIBSVM is software developed by Chih-Chung chang and Chih-Jen Lin was used for determining the value of two parameters[C, γ]. Our goal is to identify good [C, γ] so that classifier can be easily predict unknown data [i.e. testing data]. [7] LIBSVM is integrated software for Support Vector Classification, [C-SVC, nu-SVC]. It supports multiclass Classification, probability estimates, and parameter selection [6]. It provides a parameter selection tool using RBF kernel which is cross validation via grid search. A grid search had been performed on C and Gamma using an inbuilt module of libsvm tools as shown in figure 3. Pairs of C and Gamma are tried and which will be best cross validated accuracy is picked. The performance of classifiers for classes of twitter comments divided as above will be determined by measuring accuracy.

4. APPROACH

4.1 Corpus Collection

We use Twitter API for collecting Tweets from Twitter[1]. The figure below shows how the tweets are collected and step by step explanation is given in the whole algorithm included further in paper.

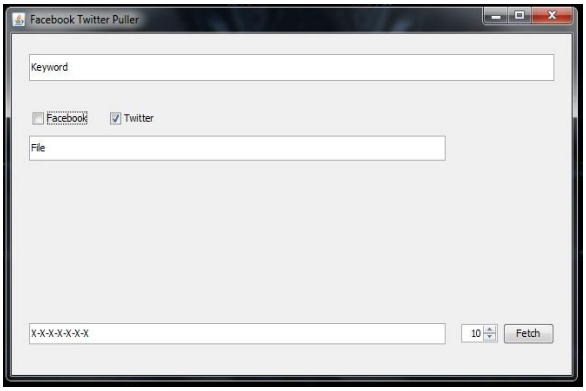

Figure 1: Twitter Tweets puller

As we can see in above figure we can fetch out tweets by clicking on fetch button as per keyword would have entered. We can fetch number of tweets we want as per requirement but there is limitation in Twitter API that it could able to extract 500 random tweets at a time. Twitter puller extract tweets from site that further will store into text file which can be then used for our purpose of opinion mining. Our tool had been developed in a way which can also able to extract tweets from twitter using Twitter API. This functionality of tool had been designed by keeping in concern that our current research work would be extended further.

4.2 Creating Domain Dictionary

We have developed a domain dictionary that contains sets of all the feature terms extracted from those tweets through the Twitter. Then the extracted feature is then enabling us to map and classify into the individual text file as shown below in figure 2. The classifier then classifies these features into three classes defined as Positive, Negative and Neutral automatically and generating files separately for each feature category respectively as shown in figure. These files generated has been strictly follow particular format supported by our training and testing tool LIBSVM and containing threshold (occurrence of word indicating opinion in tweets) of words and their synonym containing in comment. This time we perform evaluation on the basis of some specific synonyms. How this whole work get done will show in further algorithm in 3.4. This pseudo code explains whole concept and approach hidden behind Twitter Tweets collection, feature extraction and classification.

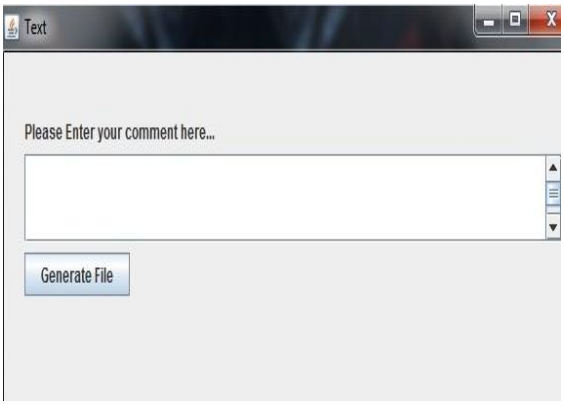

Figure 2: Classifier that classifies features of Twitter Tweets separately

4.3 Corpus analysis.

Now we have testing file in particular format containing occurrence of word in tweets would shown its impact as Positive, Negative and Neutral. We use tool LIBSVM for analysis the extracted feature from twitter. LIBSVM then firstly perform training on testing file shown accuracy level of our mined data. It further does prediction to perform evaluation and experiments on different values. These results will further shown in next section.

4.4 Proposed Methodology

Step 1**: Corpus collection**

The first step is to collect the number of Tweets refers instances from Twitter.

Step 2: **Domain Dictionary**

Creation of the dictionary consisting feature terms relevant to the annotated classes predefined.

Step 3: **Extraction from Tweets Puller tool**

The Tweets puller tool enables us to fetch all the tweets from the twitter when connected to the server.

Step 4: **Classification from Classifier Tool**

The next step is to classify those collected tweets into sub-classes as Positive, Negative and Neutral through the classifier tool. The classifier generally takes a single instance and then matches it with the features in domain dictionary containing some synonym of features. This mapping is done to generate the threshold frequency for each feature and automatically generate a text file of it.

Step 5: **Implementation of LIBSVM tool**

The generated text files is then processed in the LIBSVM tool that provides the accuracy rate for testing the classification which is further been trained and predict to be analyzed. The result of the training and predicting produces a contour graph shown in section 4.

Step 6: **Analyzing the results**

The final step is to analyze the results obtained from the contour graph and conclusions is drawn for the performance of the Classification.

The whole process done defined above will be concluded in following algorithm which clears the crystal picture of concept being used for our work:

5. RESULTS & DISCUSSIONS

The performance of our system to classification of features mined from twitter has been determined bytraining and predicted our cross validation files. We train our file and get following contour graph as shown below. It demonstrates feature extracted from Twitter tweets and distinguished it among three subclasses we made. The best accuracy we got is 74.8268% as shown below after cross validation.

Table 2. C and Gamma values for training set of twitter tweets with accuracies

Class	C	Gamma	Accuracy
Positive	32	0.0078125	74.6269%
Negative	32	0.0078125	68.6567%
Neutral	2	0.5	74.6269%

The tabulated value of C and Gamma for predicting different classes of features of twitter comments and for training dataset in given Table 2. Further, variation of C and Gamma values could provide more accuracy of training set. SVM models have a cost parameter, *C*, that controls the tradeoff between allowing training errors and forcing rigid margins. The γ parameter of the Gaussian kernel and the degree of polynomial kernel determine the flexibility of the resulting SVM in fitting the data. On using the RBF kernel with value of parameters[C= 8, γ= 0.0078125] an accuracy of 74% was obtained in distinguishing various tweets features classes from other two classes. The average accuracy of three classes is 70.592%. This proved that opinion and views posted on twitter contain impact of which could be categorized into three classes. The development of such concept will provide efficient method to classify all the opinions and views posted on twitter from different user. It will be further useful for analyzing comments and reviews that had been also found at many social websites.

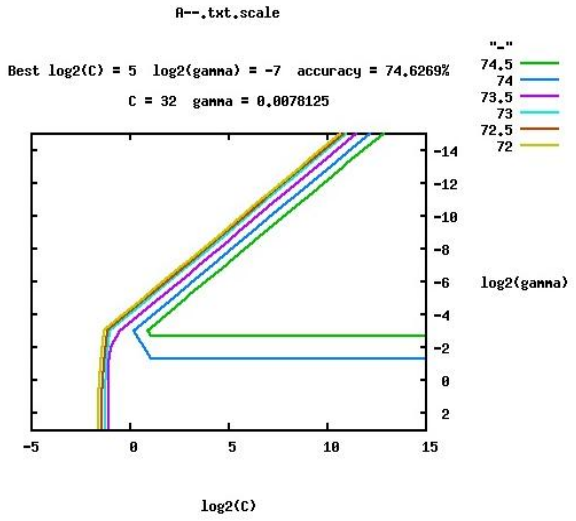

Figure 3: Shown accuracy of tested corpus of twitter

6. CONCLUSIONS

The average accuracy of 70.5% was obtained in classifying various classes. The final conclusion drawn from this research work is we have developed method which is efficient and time saving to classify millions of tweets posted on twitter. These classified opinions will then become desired data to find the reviews of users regarding any issue belong to any category. It reduces the manual survey work that had been done for

drawing conclusions on opinion posted on twitter. This work could further extended for any of frequently access social websites containing several reviews from different people.

7. REFERENCES

[1] L. Cai and T. Hofmann. Text categorization by boosting automatically extracted concepts. In SIGIR '03, pages 182.189, New York, NY, USA, 2003. ACM Press.

[2] Twitter as a Corpus for Sentiment Analysis and Opinion Mining Alexander Pak, Patrick Paroubek.

[3] Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews Kushal Dave, Steve Lawrence and David M. Pennock.

[4] D. Zhuang, B. Zhang, Q. Yang, J. Yan, Z. Chen, and Y. Chen. Ef_cient text classi_cation by weighted proximal svm. In ICDM, pages 538.545, 2005.

[5] Ivanciuc, O. Applications of Support Vector Machines in Chemistry. (2007), Rev. Comput. Chem., 23, 291-400.

[6] Chang, C.-C., & Lin, C.-J., (2003), LIBSVM: a library for support vector machines.

[7] Wei, Hsu, C., Chung Chang, C., & Chih-Jen Lin, A., (2003), Practical Guide to Support Vector Classification.

[8] Vladimir N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.

[9] Haggai Toledano; Elad Yom-Tov; Dan Pelleg; Edwin Pednault; Ramesh Natarajan, (2008), Support Vector Machine Solvers: Large-scale, Accurate, and Fast.

[10] Stephen dann, (2010), twitter content classification.

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135.

[13] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pages 275–278, Washington, DC, USA. IEEE Computer Society

[14] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In ACL, the Association for Computer Linguistics.