Opinion Extraction and Classification of Real Time Facebook Status

By Akash Shrivatava & Bhasker Pant

Graphic Era University, Dehradun

Abstract - Social media like Facebook today are not only just a website. They are now become much popular communication tool for internet users. It is a medium through which users belonging to any of category, profession can make their comments. These all comments have contained some features along with it. These comments or status are really useful which are actually viewed as their 'OPINIONS'. Opinions are really important while we need to analyze any of product, topic, discussion and whatever which will require some user opinions to draw some inferences and conclusions from them. Social media plays an important role for this intention. In this paper we focused on facebook statuses, which we can view as opinions of users or their reaction on concern we want to analyze. We develop tool status puller that automatically collects random facebook statuses. Then we make classifier that performs classifications on that corpus collected from facebook. Our classifier is able to extract three features GOOD, BAD and AVERGAE from that statuses respectively. As per classifier results we perform evaluations experiments which further can be work for feature mining of user opinions on facebook. It's pure new and unique technique proposed in the field of opinion mining.

Keywords : *Opinion mining, classification, facebook status mining, Data mining, web mining, text categorization, support vector machine.*

GJCST Classification: *H.3.5*

Strictly as per the compliance and regulations of:

Opinion Extraction and Classification of Real Time Facebook Status

Akash Shrivatava [α] & Bhasker Pant [σ]

Abstract - Social media like Facebook today are not only just a website. They are now become much popular communication tool for internet users. It is a medium through which users belonging to any of category, profession can make their comments. These all comments have contained some features along with it. These comments or status are really useful which are actually viewed as their '*OPINIONS*'. Opinions are really important while we need to analyze any of product, topic, discussion and whatever which will require some user opinions to draw some inferences and conclusions from them. Social media plays an important role for this intention. In this paper we focused on facebook statuses, which we can view as opinions of users or their reaction on concern we want to analyze. We develop tool status puller that automatically collects random facebook statuses. Then we make classifier that performs classifications on that corpus collected from facebook. Our classifier is able to extract three features GOOD, BAD and AVERGAE from that statuses respectively. As per classifier results we perform evaluations experiments which further can be work for feature mining of user opinions on facebook. It's pure new and unique technique proposed in the field of opinion mining.

Keywords : Opinion mining, classification, facebook status mining, Data mining, web mining, text categorization, support vector machine.

I. Introduction

The dramatic and exponential growth of content available on web and its classification has now become an efficient methodology to make the contents of large repository in an organized manner [1, 4]. Social networking websites are the new era of expressing views. Today every fifth person put their opinions, views, comments on these micro-blogging and social sites like TWITTER[1], FACEBOOK[2] and many more. The format and pattern include in these websites are so easy to use and this is the most genuine reason that their accessing rate exponentially increased from last few years. Authors of those comments, views and opinions write their point of perception on any of discussion topic. It may include any political issue, religious issue, technology, product, movie review and much more daily gossiping issues flooded in their surroundings [2]. Now people are using internet as a communication tool among their social network including friends, family, friends of friends. It signify that they all now moved from traditional trends like mail, blog to these micro-blogging and social network sites. But they do not even realize that by gradually putting and sharing their opinions among their friends on these sites will finally become huge and relevant repository for any of particular entity or organization. Such dataset collected from all these sites can be efficiently used for marketing, case study and social studies. Organizations that required can easily draw inferences and conclusions regarding their product, technology or political point whatever they all are concerning with by going through opinions comes from these sites [3]. It indicates that now to analyze any feedback for anything you are concerning with, there is no major need to survey it home to home or person to person individually by contacting them through any means. In spite of this just need to collect opinions from these social networking sites and draw conclusions that what people like/dislike, what are their intentions towards any issue? Likewise, many queries can be answer by analyzing just their opinions on different aspects of their life posted on these sites. We use the dataset collected from FACEBOOK. FACEBOOK contains large number of comments concerning their personal thoughts and public views from different users belonging different regions and countries. TABLE 1 shows typical example of some FACEBOOK comments. 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 requirements, demands from the current scenario. We show how to use FACEBOOK as a medium for opinion mining. We use facebook for following reason:

- FACEBOOK is well known and frequently accessing site across the globe.
- FACEBOOK is not biased to any particular people category the crowd we will get on facebook is belonging to general public whose opinions are really worthwhile for any general survey.
- FACEBOOK joined by many people from different countries belonging to different category having many languages.

We collected around 2000 comments from facebook which evenly split automatically into three sets as follows:

1. Comments containing positive impact such as Good, Best, Happy and its more synonyms collected into Good.txt file.

Author α σ : Graphic Era University, Dehradun.
E-mail α : akash.10may@gmail.com
E-mail σ : pantbhaskar2@gmail.com

2. Comments containing negative impact Bad, Worst, Sorrow and its more synonyms collected into Bad.txt file.
3. Comments containing average impact Neutral, Average, Fine and its more synonyms collected into Avg.txt file.

We show how to classify these features based on different impact through classifier that extracts features in three separate classes. Finally we use LIBSVM providing multi-classification [9] support vector machine tool to train and testing accuracy of system that up to which extent our system does opinion mining.

matthew 24:14 this good news of the kingdom will be preached in all the inhabited earth for a witness to all the nations;and then the end will come.
Had the best margharita EVER. you know its good when you have a slight burning sensation in your throat.
Nursing, hockey, and some quality time with dad...today life is amazing. Hopefully it keeps running into tomorrow when I finally get some quality time with an awesome friend!
This will teach those pompous pricks to get their hoity toity higher educations! Except athletes: they're good hardworking people who deserve special breaks.
I made an 84 on my math test and my average is an 88!!!! Whoot whoot yes im freakin excited!

Table 1: Example of Facebook Status with User Views

a) *Contribution*

The contribution of our paper is as follows.
1. Our method shows that how feature can be extracted from comments posted on FACEBOOK on the basis of which inferences can be drawn according to requirement.
2. We have a Facebook status puller which can collect 500 facebook comments at a time. No human efforts need to collect corpus. It is as flexible as according to desire user can collect corpus as per keywords on facebook.
3. We develop a classifier that classify collected corpus from facebook into three classifications which would automatically store as per their feature in separate files. It again reduces time and effort.

[1]http://twitter.com
[2]http://facebook.com

4. After collecting corpus we can do linguistic analysis on that corpus.
5. We can also build sentiment classification system based on features including in comments.

We conduct experimental evaluations to produce real time results on a set of real facebook comments posted to prove that our technique is efficient enough and performs better than previously proposed methods.

b) *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 extraction facebook comments, training and testing data. In section 3, we give the explanation of approach for collecting the corpora and its classification. Furthe experimental evaluations performed by LIBSVM shown in section 4. Finally we conclude our paper about our work.

II. Material and Tool Used

a) *Data Used*

Facebook comments are used for our research work which is our primary focus. They will be further use for mine opinion on the basis of features contain in the comments extracted.

b) *Support Vector Machine*

Support vector machine is kernel based techniques which is major development in the machine learning algorithms. Support vector machines are groups of supervised learning that can be efficiently apply for classification. It represents an extension version to non linear model generalized portrait algorithm developed by Vladimir Vapnik [8]. The algorithm adopted in SVM is based on the statistical learning theory and the Vapnik-Chervonenkis [VC] dimension introduced by Vladimir Vapnik and Alexey Chervonenkis. A support vector machine [SVM] does classification as by constructing N-dimension hyperplane that optimally divided the data into two categories. [5] Even without feature selection performance of SVM can be very efficient [10].

c) *SVM Implementation- LIBSVM*

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 [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 facebook comments divided as above will be determined by measuring accuracy. SVM is known to be the most

keyword in our developed tool. How our tool collect data from facebook shown in figure below and explain step by step in the whole algorithm included further in paper.

III. APPROACH

a) Corpus Collection

We use Facebook API for collecting facebook comments from facebook1. We queried facebook as per
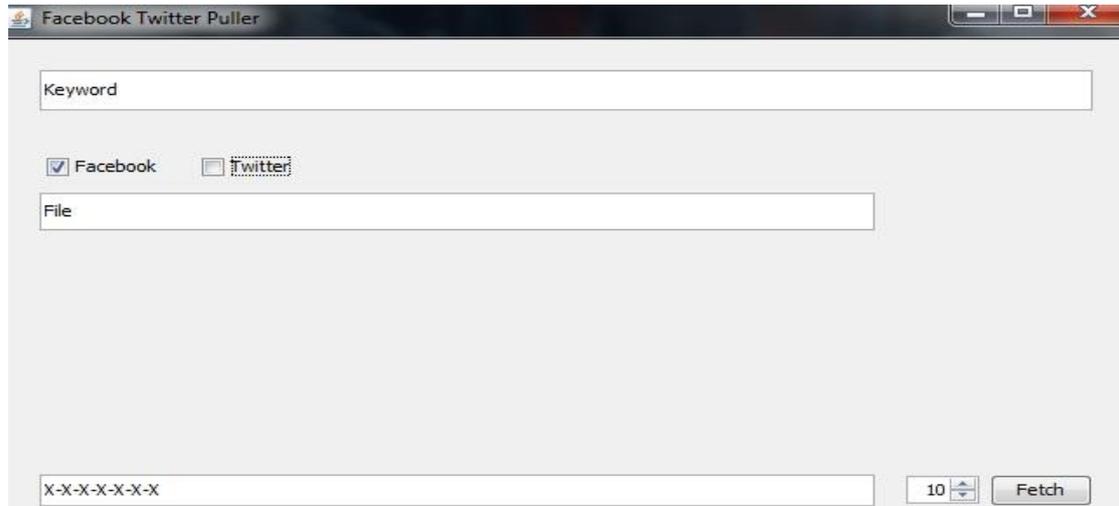
Figure 1: Facebook status puller

As we can see in above figure we can fetch out comments by clicking on fetch button as per keyword would have entered. We can fetch number of comments we want as per requirement but there islimitation in facebook API that it could able to extract 500 random comments at a time. Facebook puller extract comments 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.

b) Feature Extraction and Classification

We collected facebook comments above, which further undergone for feature extraction from those comments individually through classifier we developed

as shown below in figure 2. This classifier then classifies these features into three classes defined above 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 comment) of words and their synonym containing in comment. The synonym of particular category which defines for our research work can be further extending for more refine research. This time we perform evaluation on the basis of some specific synonym. How this whole work get done will show in further algorithm in 3.4. This pseudo code explains whole concept and approach hidden behind facebook comments collection, feature extraction and classification.
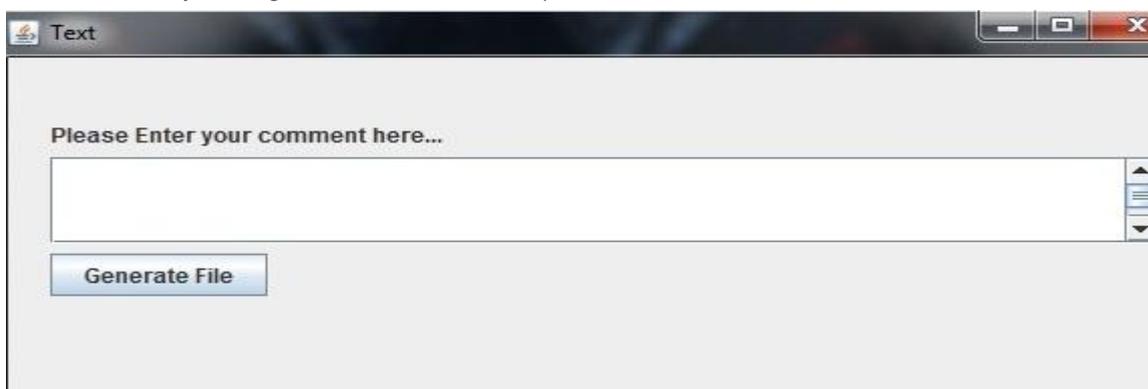
Figure 2 : Classifier that classifies features of facebook comments separately

c) Corpus Analysis

Now we have testing file in particular format containing occurrence of word in facebook comment would shown its impact as good, bad and average. We use tool LIBSVM for analysis the extracted feature from facebook comments. 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.

d) Proposed Methodology

Step 1: Corpus collection

The first step is to collect the number of comments refers instances from Facebook.

Step 2 : Extraction from Status Puller tool

In this Step the real-time comments from the Facebook status is been pulled from the status puller tool when connected to the server.

Step 3 : Classification from Classifier Tool

The next step is to classify those collected comments into sub-classes as Good, Bad and Average 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 4 : Processing 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 traine and predict to be analyzed. The result of the training and predicting produces a conture graph shown in section 4.

Step 5 : Analyzing the results

The final step is to analyze the results obtained from the conture 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:

IV. Results and Discussions

The performance of our system to classification of features mined from facebook comments has been determined by training and predicted our cross validation files. We train our file and get following conture graph as shown below. It demonstrates feature extracted from facebook comments and distinguished it among three subclasses we made. The best accuracy we got is 74.8268% as shown below after cross validation. The tabulated value of C and Gamma for predicting different classes of features of facebook comments and for training dataset in given Table 2.
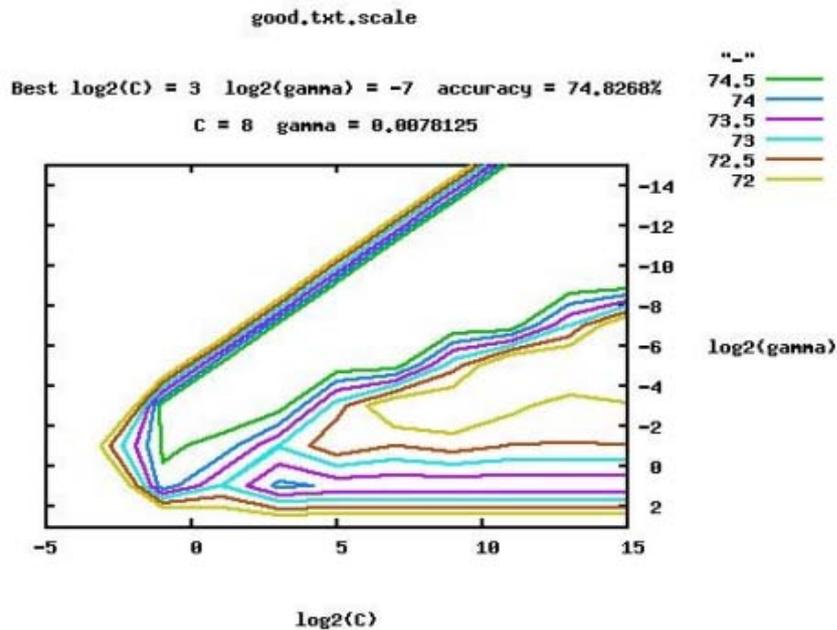

Figure 3 : Shown accuracy of tested corpus of facebook

Class	C	Gamma	Accuracy
Good	8	0.0078125	74.8268%
Bad	0.5	0.5	69.515%
Average	2	0.5	67.4365%

Table 2 : C and Gamma values for training set of facebook comments with accuracies

Further, variation of C and Gamma values could provide more accuracy of training set. On using the RBF kernel with value of parameters[C= 8, γ = 0.0078125] an accuracy of 74% was obtained idistinguishing facebook comments features classes from other two classes. The average accuracy of three classes is 70.592%. This proved that opinion posted on facebook contain impact of view 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 facebook from different user. It will be further useful for analyzing comments and reviews that had been also found at many social websites.

V.   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 very efficient and time saving method to classify millions of comments posted on facebook. These classified opinions will then become required data to judge the reviews of users regarding any concern belong to any issue. It reduces the manual survey work that had been done for drawing conclusions on opinion posted on facebook. This work could further extended for twitter tweets or any of frequently access social websites containing several reviews from different people.

References Références Referencias

1. L. Cai and T. Hofmann. *Text categorization by boosting automatically extracted concepts*. In *SIGIR '03*, pages 182.189, New York, NY, USA, ACM Press. (2003).
2. Alexander Pak, Patrick Paroubek. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* In Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10 Valletta, Malta: European Language Resources Association ELRA (May 2010).
3. Dave, Steve Lawrence and David M. Pennock. *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews Kushal* (2003).
4. D. Zhuang, B. Zhang, Q. Yang, J. Yan, Z. Chen, and Y. Chen. *Efficient text classification by weighted proximal SVM.* In *ICDM*, pages 538.545, (2005). 5) Ivanciuc, O. *Applications of Support Vector Machines in Chemistry.* Rev. Comput. Chem., 23, 291-400, (2007).
5. Chang, C.-C., & Lin, C.-J., *LIBSVM: a library for support vector machines* (2003).
6. Wei, Hsu, C., Chung Chang, C., & Chih-Jen Lin, A. *Practical Guide to Support Vector Classification.* (2003).
7. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, (1995).
8. K. Crammer and Y. Singer. *On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res.*, 2:265.292, (2002).
9. H. Taira and M. Haruno. *Feature selection in svm text categorization.* In *AAAI '99/IAAI '99*, pages 480.486, Menlo Park, CA, USA, (1999).

This page is intentionally left blank