The Opinion Mining from Social Media by using Support Vector Machine (SVM) Algorithm

Thiri Yadanar, Htun Htun, Nyein Chan Soe

Abstract - **The proliferation of social media websites allows users to communicate with each other by using several tools like chats, discussion forums, comments etc. At any type, any category and any word user professionally can make their comments. These all comments have some features and attributes with it. These comments or status are really useful which are actually viewed as their 'OPINIONS'. Opinions are really important when we are analysing any of product, topic and discussion. Then, we introduce and present our original method of opinion classification and we test the presented algorithm on real world datasets, reporting on the results. The aim of our research is to extract three features POSITIVE, NEGATIVE and NEUTRAL from that statuses respectively. It performs evaluations experiments for each classifier results which can be worked for feature mining of user opinions on Social media. Then, after we presented our original method of opinion classification and we test the real-world datasets, reporting on the results. Analysing the encouraging results of our study, exploring the suitability and performance on Support Vector Machine (SVM), will also be presented.**

Index Terms - **Opinion Mining, Social media, Support Vector Machine, Text Analysis**

1). INTRODUCTION

Social networks have become widely-used and popular mediums for information technology and political, social obligation, economical works and other as well as functions of social interactions. The one point of important is users' contributions and activities provide a valuable insight into individual behaviour, experiences, opinions and interests. Considering that personality, mental processes and affective reactions of users or human, there is an excessive opportunity for adding new personality-based qualities to user interfaces. Personalized systems used in domains such as, e-learning, information filtering, collaboration and e-commerce could greatly benefit from a user interface that adapts the interaction (e.g., motivational strategies, presentation styles, interaction modalities and recommendations) according to user's personality. Having captured past user interactions is only a starting point in explaining the user behavior from a personality point of view.

Social networking websites can help to close each other and eliminate coordination issues among individuals that are at a considerable physical distance. It can build the opportunity of social campaigns by collecting and extracting the required data at any place and at any time. In any case, in social networking sites, individuals for the most part use unstructured or semi-structured language for correspondence. People mean to pay less attention to spellings, grammar test and precise linguistic development of a sentence in many discussions, posts and comments.

Useful of text mining techniques on social networking fields can discover important results of outputs about communication practices between people.

This research builds upon previous inter-disciplinary research works regarding personality as it pertains to the design of intelligent interactive systems. The new communication technologies have abilities that more information to consider, though the process of their utilization is far from straightforward. Intelligent technologies are expected to play a prominent role in bringing these data to a new level of usability. A variety of Facebook variables were expected to analyze an obvious role in establishing appropriate context for our particular investigations. Facebook profiles and activities provide valuable indicators of user's personality, explain the actual, rather than idealized or projected personality.

In the summary, this system shows how to classify these features based on different impact through classifier that extracts features in three separate classes. We use SVM providing multi-classification support vector machine tool to train and testing accuracy of system that up to which extent our system does opinion mining.

This paper organizes as following. After introducing with section 1, in section 2, it discusses the related works with the paper. In section 3, it describes background theory and section 4 presents the experiment of the system. And section 5 is allowed the evaluation result then finally section 6 shows conclusion for this paper.

2). RELATED WORKS

The correlation between users' social network activity and personality has been presented by the focus of several studies in the last decade (Bai, Zhu, and Cheng 2012; Golbeck, Robles, and Turner 2011; Bachrach et al. 2012). Personality traits of the Chinese most popular social network RenRen users were analyzed in (Bai, Zhu, and Cheng 2012). C4.5 Decision Trees have shown the best results, yielding 69-72 percent accuracy, for a combination of features related to

Manuscript received September, 2018.

Thiri Yadanar, *is with the Department of Information Technology of West Yangon Technological University, Yangon, Myanmar (Corresponding author to provide phone: 092015043).*

Dr. Htun Htun, *is with the Department of Information Technology of West Yangon Technological University, Yangon, Myanmar.*

Nyein Chan Soe, *is with the Department of GIS of University of Computer Studies, Yangon, Myanmar*

users' network activity along with affective linguistic features extracted from statuses and blog posts.

The work most closely related to our own is (Golbeck, Robles, and Turner 2011). Two regression techniques, namely m5sup/Rules and Gaussian Processes, were applied to build predictive personality models. The authors consider users' Facebook data through parameters such as structural characteristics, personal info, activities and preference, in addition to the linguistic attributes extracted with LIWC from the users' statuses. The lack of demographic diversity in participant sampling was one of the major drawbacks for generalizing the results of the last two studies, Chinese population and authors' Facebook friends respectively. Few studies using considerably larger number of instances from the same dataset under our investigation have a rather different objective from ours, namely to examine the correlations between the personality traits and Facebook activity data (Bachrach et al. 2012) and the associations between personal attributes and Facebook Likes (Kosinski, Stillwell, and Graepel 2013). These studies were not meant to look at the rich linguistic patterns that occur in the language use on social networks, which is in the focus of this research.

The 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. 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 in [5].

3). Background Theory

1). 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 applied for classification. It represents an extension version to nonlinear model generalized portrait algorithm developed by Vladimir Vapnik. 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. Even without feature selection performance of SVM can be very efficient. The architecture of the SVM is shown in Figure 1.

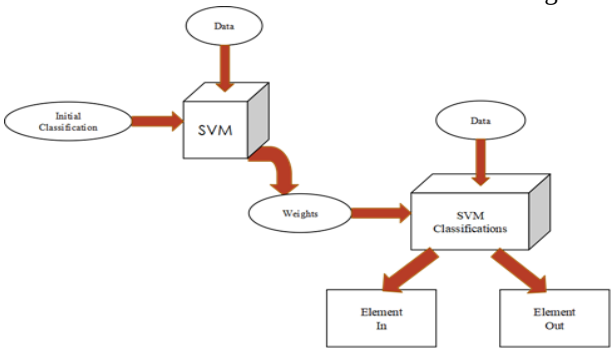

Figure 1. Architecture of the Support Vector Machine

2). Text Analysis (Opinion Mining)

Opinion mining is research subtopic of data mining aiming to automatically obtain such useful knowledge. It has been widely used in real world applications such as ecommerce, business-intelligence, information monitoring and public pull. It seeks to determine the sentiment, attitude or opinion of an author expressed in texts with respect to a certain topic. On the Facebook, there are increasing numbers of pages, where user posts their comments and provides their opinion like good, bad or average. These pages are important resource providing advice to new users and helping them.

A. Text Mining Vs. Data Mining

As compared to data mining, there is more convolutions in text mining because of its unstructured and obscure content. Text mining contains features of data mining, but the distinctive point between these processes is that data mining tools are designed to cope with structured data from databases, while text mining is able to handle unstructured or semi-structured data sets which include full-text documents, emails, and HTML files etc. Consequently, text mining has more promising results for its users. Hitherto, data mining (using structured data) has attracted the attention of researchers. The problem affiliated with text mining is evident (i.e., the incapability of computers to comprehend natural language). The main goal of text mining is to explore the unknown information that is concealed from masses. The handling and mining process of data is significant, as it is the raw form of information, which eventually ends up by creating knowledge.

B. Association Rule Extraction

Association rule extraction proposed in the system that keeps significance in text mining field. It involves finding association relationship between different feature words from the text collection. Such exploration of interesting association relationships among huge amounts of transaction records will facilitate in many decision-making processes.

3). Dependency Relation for Feature Extraction

Dependency grammars represent sentence structures as a set of dependency relationships. A dependency relationship is an asymmetric binary relationship between a word called head or governor, and another word called modifier or dependency. The dependency of words will form a dependency tree using POS tagger. From this tree, we attempt to capture the relation product feature and opinion using dependency relations between them. The structure of a sentence consists of dependencies shown in Figure 2.

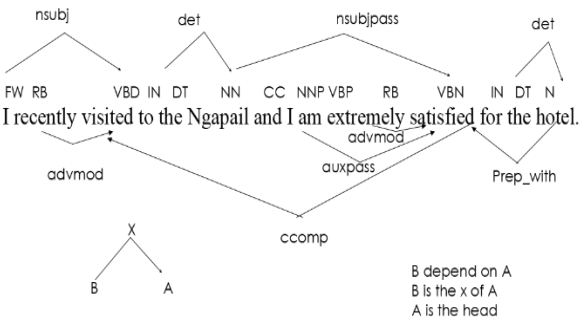

Figure 2. Dependency Relationship of the Sentence

A. Personality Data

The study of personality reflected in user's Facebook activities includes a wide range of features. Some of the most intuitively predicted indices are the statistical data for user's activities (e.g., number of likes, statuses, groups, tags, events). Demographic characteristics such as: age and gender, were accounted for since their effect is known to manifest in the context under investigation (Golbeck, Robles and Turner 2011). Egocentric network parameters representing the number of friends, and measures such as density, brokerage, and provide additional insight into user's social behavior instrumental in assessing personality along several dimensions (Celli et al. 2013).

4). Standford parser

In preprocessing step, each comment is parsed using Standford parser, which provides POS tags to English words based on the context in which they appear. And then the comments sentence is converted into dependency relations using Standford parser. The dependency relations encoded the grammatical relations between every pairs of words. For each segment, the system extracts feature that correspond to the context by the pattern-based extraction. Following figure 3 is shown to analyze the features of the comments by using Standford parser in experiment.

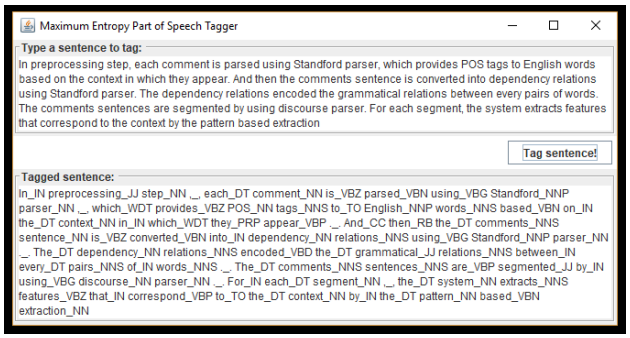

Figure 3 Analyzing the features of the comments by using Standford parser

A. Extraction feature Words

Feature extraction words are extracted in this phase. Since, to know how intense the features for the customer is also important for the decision-making process, we need to extract the negation and modifier words. The linguistic filtering pattern to identify the opinion phases is the following Single verb such as "like", "satisfy", and so on Single adjective such as "good", "great" and so on. One or more adjective such as "nice little", "not good". one or more adverb and an adjective such as "very good", "not so bad". The system information is used to classify service feature opinion pair. To reduce the variation of linguistic constructions, assume that the shortest dependency path tracing from a product feature through the dependency tree to an opinion word. This system intended to learn such pattern from the dependency paths for each relationship.

5). Apriori Algorithm

The apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules. Apriori uses "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data). Here is the apriori algorithm in pseudo code

1. Let k = 1
2. Generate frequent itemsets of length 1
3. Repeat until no new frequent itemset are identified
   a. Generate length (k+1) candidate itemsets from length k frequent itemsets
   b. Prune candidate itemsets containing subsets of length k that are infrequent
   - How many k-itemsets contained in a (k+1)-itemset?
   c. Count the support of each candidate by scanning the DB
   d. Eliminate candidates that are infrequent, leaving only those that are frequent

4). PROPOSED SYSTEM ARCHITECTURE

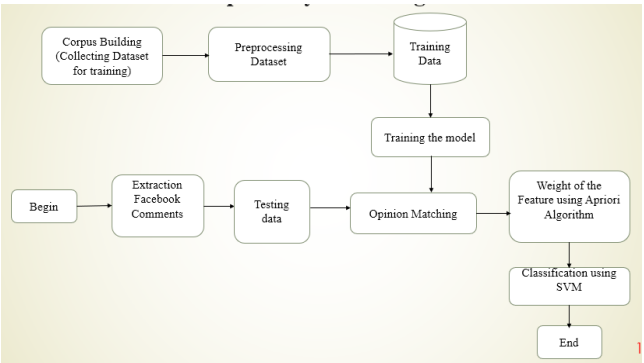

Figure 4. Proposed System Architecture

1). Proposed Methodology

There are five steps in the proposed system:

Step 1: Corpus collection: It is to collect the number of comments refers instances from Facebook.

Step 2: Pulling up: This pulls the real-time comments from the Facebook status from the status puller tool when connected to the server.

Step 3: Classification: Classifier Tool classify those collected comments into sub-classes as Positive, Negative and Neutral.

Step 4: Processing of SVM: These algorithms generate text files and then processed in these algorithms that provides the accuracy rate for testing the classification which is further been trained and predict to be analyzed.

Step 5: Analyzing the results: It is to obtain the results from the construe graph and conclusions is drawn for the performance of the classification.

In experiment, we use Facebook API for collecting Facebook comments from Facebook. We collected Facebook comments, which further undergone for feature extraction from those comments individually through classifier we developed as shown below in figure 5.

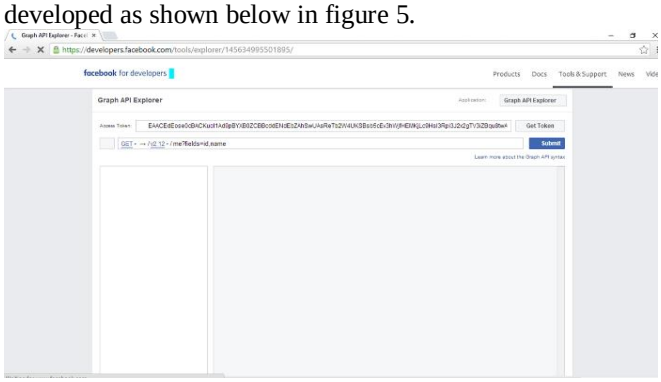

Figure 5: Extraction comments from Facebook API

After that the sample of 250 user instances from Facebook activity and demographic data) with approximately 10,000 status updates used in our system. We use Facebook API for collecting Facebook comments from Facebook pages. We required Facebook as per keyword in Facebook pagers Tools. We will collect the comments from the Facebook pages and stored in CSV form. Using POS tagging for feature extraction. The weight of the features is also calculated by using Apriori algorithm in order to discover the frequent features. Now we have testing file in particular format containing occurrence of word in Facebook comment would show its impact as good, bad and average. We use tool SVM for analysis the extracted feature from Facebook comments. SVM 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.

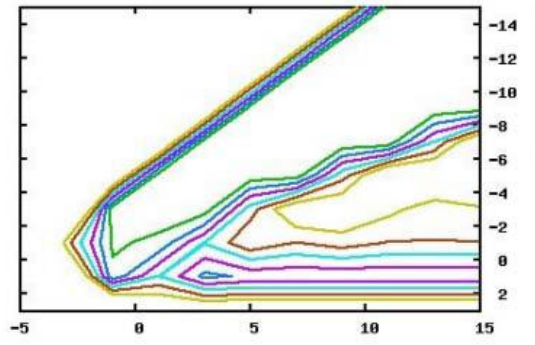

Figure 6: Accuracy of tested corpus of Facebook

5). Evaluation Result

This system has been developed very efficient and time saving method to classify millions of comments posted on Facebook. This system got the best accuracy by using SVM algorithm. The performance of our system to classification of features mined from Facebook comments are determined by training and predicted our cross-validation files. We trained our file and get following couture graph. It demonstrates feature extracted from Facebook comments and distinguished it among three subclasses we have already made**.**

1. Analysis of the result

It consists of 10775 comments from the comments posted on Trip advisor Facebook page. This part of data is not used during training period. Since the count of tokens should be in whole number, some consideration has made about the percentage of testing data to make it whole number. The test data contains total of 10775 randomly selected comments out of which 82% are good, 13% are average and 5% are bad. So, this page is most liked by many people. It is shown in the following pie chart.

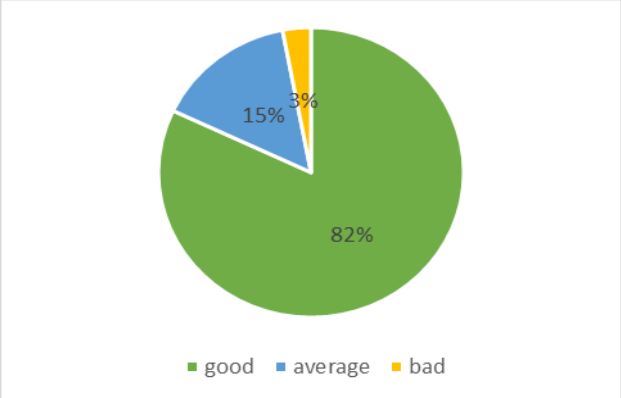

Figure 7. Analysis result of Dataset

2. Accuracy Measure of the Proposed System

We then proceeded with refining the classification mechanisms by considering better sampling of features based on the Pearson correlation coefficient. The choice to include features maximizing the correlation with a trait and minimizing the correlation with other features, has limited the number of features to 5-16, and yielded precision improvement of up to 78%. The improvement was spread across other algorithms, such as decision trees, rule-based algorithms and expected result for these algorithms are proven to perform better for selection of features with higher information gains. The size of training data is gradually increased and the performance of the algorithm is observed. The result so found is presented in the table 5.1.

Table 1. The accuracy for Different Training

Training Data Comments	10000	20000	40000	80000
SVM	71%	75%	79%	90%

So, it performs very well for the small training data size as well as for large data size.

6). Conclusion

In the system, the final conclusion will be 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.

These propose system will reduce the manual survey work that had been done for drawing conclusions on opinion posted on Facebook. This work could further have extended for twitter tweets or any of frequently access social websites containing several reviews from different people.

7). Acknowledgment

First of all, the author really thanks to Chairmen and Co-chairmen of Organization Committee for International Journal of Science, Engineering and Technology Research (IJSETR). She also like to express her supervisor, her Head of Department, and all of her teacher from Department of Information Technology, West Yangon Technological University. She has provided not only helpful guidance but also a lot of inspiration, motivation and encouragement during my research work.

Finally, she is grateful to my parents and my family who specifically offered strong moral and physical support, care and kindness, during the year of my M.E study and anybody for overall supporting during my thesis.

References

1) Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. 2012. Personality and Patterns of Facebook Usage. In Proceedings of the Third

Annual ACM Web Science Conference, 36- 44. New York, NY, USA: ACM Press.
2) Back, M.; Stopfer, J.; Vazire, S.; Gaddis, S.; Schmukle, S.; Egloff, B.; and Gosling, S. 2010. Facebook Profiles Reflect Actual Personality, Not Self-Idealization. Psychological Science 21(3): 372-374.
3) Bai, S., Zhu, T., and Cheng, L. 2012. Big-Five Personality Prediction Based on User Behaviors at Social Network Sites. Eprint arXiv:1204.4809.
4) Celli, F., Pianesi, F., Stillwell, D. S., and Kosinski, M. 2013. Workshop on Computational Personality Recognition (Shared Task). The Seventh International AAAI Conference on Weblogs and Social Media. Boston, MA, USA.
5) K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res., 2:265.292, (2002).
6) 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).
7) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. Rev. Comput. Chem., 23 291-400, (2007).
8) Evans, B. M., Kairam, S., & Pirolli, P. (2010). Do your friends make you smarter?: An analysis of social strategies in online information seeking. Information Processing & Management, 46(6), 679-692.
9) Li, J., & Khan, S. U. (2009, November). MobiSN: Semantics-based mobile ad hoc social network framework. In Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE (pp. 1-6). IEEE.
10) Sorensen, L. (2009, May). User managed trust in social networking-Comparing Facebook, MySpace and Linkedin. In Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronics Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on (pp. 427-431). IEEE.

Thiri Yadanar received her BE (Information Technology) degree from Technological University (Mandalay), Mandalay, Myanmar. She is doing postgraduate research for master degree at Information Technology Department, West Yangon Technological University. Her research is concerned data mining technology. She is also a lecturer at Pyay Technology University.

Dr. Htun Htun (Associate Professor), Department of Information Technology of West Yangon Technological University, Yangon, Myanmar.

Nyein Chan Soe (Ph.D Candidate), Department of GIS of University of Computer Studies, Yangon, Myanmar.