Machine Learning Based Accurate Prediction of Opinion Polarity Potential for Improvement and Optimization

Chanchal Soni
M.tech Scholar
Department of Computer Science Engineering
Maharana Pratap College of Technology, Gwalior
chanchal147soni@gmail.com

Unmukh Datta
Associate Professor
Department of Computer Science Engineering
Maharana Pratap College of Technology, Gwalior
unmukh62@hotmail.com

Abstract-Opinion Mining is the computational detailed investigation of people's attitudes, opinions, and emotions concerning of events, topics or individuals. The micro-blogging and social network sites are considered an extremely best source of information for the reason that people discuss and share their sentiments about a certain subject freely. Machine learning is about forecasting the future based on the past information. SVM is a supervised machine learning procedure which can be applied for classification. We propose a novel method for semantic analysis and opinion mining to solve different sentiment polarity. The method provides automatically preprocessing of data and extract words from a sentence using machine learning. The proposed algorithm used improved Support Vector Machine method for better opinion mining and accurate prediction. The experimental outcome demonstrate that system is well suitable for opinion polarity.

Keywords: Opinion Mining, Sentiment Analysis, SVM, Twitter, Polarity

I. INTRODUCTION

Twitter, with more than 350 million monthly active users and over 520 million tweets per day. Twitter has nowadays become a goldmine for individuals and organizations who have a strong political, social, or economic concern in enhancing and maintaining their reputation and clout and. Opinion mining[1] provides these organizations the capability to monitor dissimilar social media web sites in real time. Opinion mining is the process of automatically identifying whether a post segment contains opinionated or emotional content, and it can likewise determine the post's polarity. Opinion mining classification aims to categorize the opinion polarity of a tweet as negative, positive, or neutral. Posts are generally composed of poorly structured, incomplete, and noisy sentences, irregular expressions, non-dictionary terms, and ill-formed words. Preprocessing means removing URLs, removing stop words, and replacing negations from users post. A series of pre-processing are applied to reduce the amount of noise in the posts before feature selection. Pre-processing is accomplished comprehensively in existing methodologies, specifically in machine learning-based methods.

Opinion Mining (OM) also termed as Sentiment Analysis (SA) is the computational analysis of public's attitudes, opinions, and sentiments or opinions concerning an entity. The object can signify events, individuals, or topics. These subjects are furthermost likely to be concealed by analyses. The two expressions OM or SA are interchangeable and express a common meaning. Opinion Mining is an unending field of research in document mining field. Opinion mining will review different post of users and mine their opinion about related subjects.The Clustering and natural language processing procedure will be applied for opinion mining.

A portion of opinion mining denotes using of natural language processing (NLP)[2] by suggested dissimilar method of dictionary for sentimentality analysis of text data as lexicon, corpus, and specific language dictionary

With the rapid expansion of company or organization have more services and products online and enhance customer satisfaction. The provider will read customer review and other customers who need to use services or products will read review to express opinions on the services. The number of customer review is increasing or huge from website, blogs, forums and social media, which the services or product is interesting. Therefore, many customers will read comment randomly which is hard to read all comments and make decision the services or products. If customer reads a few reviews, customer might get opinion review to be bias. Therefore, opinion mining is a technique of field area of information extraction from text processing, which is benefit and many opportunities to improve or develop factor to business work by this analysis. Analysis of sentence has level of sentiment from emotion word and calculated score of similarity or cluster with the kind of word as positive or negative called sentiment polarity[4].

In many applications, it is important to consider the context of the text and the user preferences. Using TL techniques, we can use related data to the domain in question as a training data. The natural language processing tools can be used to facilitate the OM process. It gives better natural language understanding and thus can help produce more accurate results of OM. That is why OM need to create more research on context-based OM. Using NLP software to strengthen the OM process has fascinated academicians recently and still necessities of some enhancements.

The rest of the paper is organized as follows. Section 2 concentrates on the literature survey. Section 3 provides the proposed steps and algorithm. Section 4 provides the implementation and result analysis. Finally, Section 5 provides concluding remarks.

II. LITERATURE SURVEY

The three main classification levels in OM are: aspect-level, document-level, and sentence-level OM. Document-level OM[3] objective to categorize an opinion document as expressing a negative or positive opinion or sentiment. It deliberates the complete document a basic information component. One component means talking about one topic. Sentence-level OM intentions to classify opinion expressed in for each sentence. The primary step is to recognize whether the opinion is objective or subjective. If the opinion is subjective, Sentence-level OM will decide whether the sentence states negative or positive opinions.

The datasets used in OM are an essential problem in opinion review field. The foremost main sources of data are from the consumer product reviews. The people's reviews are significant to the industry holders as they are takings business decisions agreeing to the analysis outcomes of users' thoughts about their products. The assessments sources are primarily review sites. OM can similarly be applicable on news articles,political debates or stock markets. For example in political debates, the investigators could figure out public's opinions on a certain political parties or election candidates. The election outcomes can also be forecast from user's political posts. The micro-blogging sites and social network sites are considered a good source of customer information because people discuss and share their opinions nearly a certain topic freely. They are similarly used as data sources in the OM process.

At the present time, a company or organization make available a business service which essentials to get feedback from consumer. The consumer review is significant to progress service for organization or company, which have both close opinion and open opinion. The open opinion means the comment as text which shows emotion and comment directly from customer. However, the company has many contents or group to evaluation themselves by rating and total rating for a type of services which there are many customer who needs to review.

[1] proposes the analysis and prediction rating from customer reviews who commented as open opinion using probability's classifier model. The classifier models are used case study of customer review's hotel in open comments for training data to classify comments as positive or negative called opinion mining. In addition, this classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques, which gives correctly classifier to 94.37% compared with decision tree Techniques. The proposed methodology used Thai customer review's hotels from a website of hotel agent service, which service in hotel reservation directly. The target of classify customer review from this website because the comment is posted from customer who is serviced checked-in and checked-out from hotel. The process is started from collected data and preprocessing is cleaned data by removal stop words and using the high frequency of word which will be selected into attribute for using classifier model. The classifier model will be solve the text of customer review that is positive of negative from training data and test data which are train from behavior posting from customer of hotel service group.

A part of opinion mining refers using of natural language processing (NLP) by proposed different method of dictionary for sentiment analysis of text as corpus, lexicon and specific language dictionary [5]. They tried to extract word from sentences for removal stop word or unnecessary word automatically. In addition, various dictionaries are solved by machine learning methods [6][7], which try to rank scoring of various dictionaries. For example, the paper in [8] used fuzzy logic algorithm to collect the ranking of different dictionary into rule for classify the opinion.

After word segmentation process is removal stop words by dictionary checking. The group of researches in [9] focuses on the calculating polarity of words to trend in positive or negative in a cluster of interest's customer that are extracted from texts and compared the word occurrence of whole sentence. If the word extractions have weight from dictionary of emotional words, it is calculated to answer the comment as positive or negative.

However, the customer review has different behavior with the product. The proposed classifier model is presented using association rule in [10][11]. The popular classifier model is naYve bayes compared with other model [12][13][14], which there are different sources such as social media and web site. From these researches are used classifier models that are the same objective to classified opinion. Our approach is different from them, this paper use the advantage of classifier model to generate the rating value from classifier which is not only shown classify opinion as positive and negative and also factors analysis[15] to impact the customer who posted or commented to positive and negative.

III. PROPOSED WORK

Proposed step

Opinion mining steps

1. Social media data collection
2. Opinion reviews
3. Opinion identification
4. Opinionative words and phrases
5. Feature selection and extraction
6. Opinion classification
7. Opinion polarity

1. Social media data collection: The first step in opinion mining is to collect large amount of data from social media like Facebook, and Twitter.

2. Opinion review: The second step is to review all the opinion from collected data.

3. Opinion Identification: The next step is to identify the opinion of the users.

4. Opinionative words and phrases: In this step we opinionative all the words and phrases.

5. Feature selection: The next step is feature selection. In opinion classification opinion analysis task is considered an opinion classification problem. The first step in the opinion

classification problem is to select and extract text features. Some of the existing features are terms frequency and presence. These features are distinct word n-grams or words and their frequency sum total. It either gives the words binary weighting. In binary weighting zero if the word appears, or one if otherwise. It also uses term frequency weights to indicate the relative importance of features.

Parts of speech: In parts of speech process finding adjectives, as they are significant indicators of opinions polarity.

Opinion phrases and word: These are words generally used to express opinions comprising bad, or good or hate or like.

Negations: The presence of negative words could change the opinion meaning like not good is comparable to bad.

6. Features: The next step is to extract the required features from the available features.

7. Opinion classification: The next step is to classify the opinion according to requirement.

8. Opinion polarity: The last step is to polarize the opinion.

```
Algorithm: Improved opinion clustering and polarity search algorithm
Input:
    AP: All posts in dataset
    F: Prominent keyword
    P: posts
    TP: Twitter posts
    FP: Facebook posts
Output:
    Positive opinion, negative opinion, sentiment polarity, RMSE, percentage accuracy
Procedure
Step 1: Collect post from social media like Facebook, Twitter
Step 2: Cluster the data according to P
Step 3: if P==TP then
        Apply Twitter API to the post
        Apply Google SDK and Opinion mining API

    Else
        Apply Facebook API to the post
        Apply Google SDK and Opinion mining API
    End if
Step 4: Apply preprocessing method
    Clearing the text
    Removing URL’s
    Removing Tags
    Removing irrelevant contents
Step 5: Apply feature selection and extraction

Prominent keyword extraction
AP= All posts in dataset
Initially F=NULL
For every x in AP do
    Y= Extract keyword from posts
    For ap_key in posts_keywords do
        If ap_key in p_key then
            F[ap_key]= F[ap_key] + 1
        Else
            Insert in K to pa_key
        End if
    End for
End for
For key in  AP do
    If AP[key] < threshold then
        Remove key from AP
    End if
End for
Return AP
Step 6: Apply Support vector machine algorithm for opinion classification
Step 7: Classify positive and negative opinion according to customer P
Step 8: Calculate the opinion polarity
    If key_match > 80 then
        Print positive accuracy
    Else
        Print negative accuracy
    End if
Step 9: Calculate RMSE
Step 10: Stop
```

The posts in social media are differ constructed on the opinion status linked with the contents.The total sentiment and opinion related with a post can be neutral, negative orpositive. The assessment of opinion at keywords level is made for sentiment analysis which enables posts classify based of attitude of the subject.

Primarily the user posts are preprocessed to take out stop words. The next step is keyword extraction. The keywords means the important words in a user post that refer to the subject of related post. The keyword lists linked with entire posts are combined together. The list of final keyword is expressed by filtering out everyfound keywords that are not as much of used in posts. The algorithm focusing on opinion and polarity associated with common and most keywords being used in every posts.

The dissimilarity in the opinion scores associated with the found keywords can be a key contributing feature in clustering the related posts constructed on sentiments. The procedure of extracting main prominent keywords is as termed in Algorithm which provides as outcome, the prominent words. The threshold differs dependent on the dataset.

IV. IMPLEMENTATION AND RESULT ANALYSIS

For implementation environment used i3 3.0 GHz machine with 4GB RAM. The result and performance graph is also discussed. Social media data used for opinion mining and polarity finding. The social media post is collected data from social web sites like Facebook and Twitter. The different post of users are reviewed and mine their opinion about related subjects.

For implementation Python programming language, social website datasets, google cloud SDK, Facebook and Twitter SDK, natural language processing libraries are used.

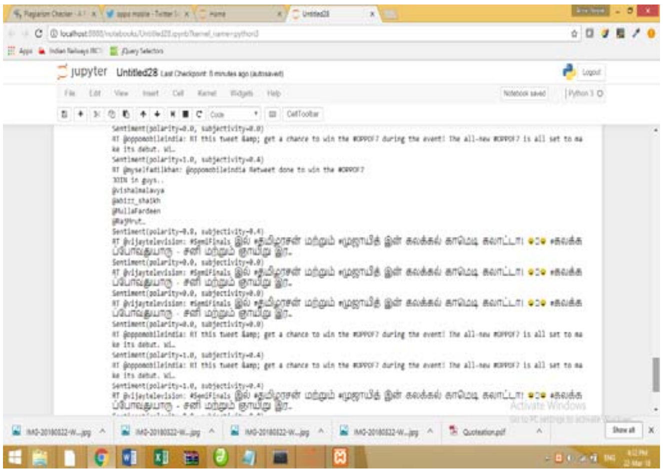
The dataset has been collected from real-world environment of online social networks. The different posts from customers related to different products are collected. Twitter posts as well as Facebook posts used for experiment. Five different groups have been chosen for performing the testing of the proposed work.

Dataset properties

Group	Members	Post
G1	2365	20321
G2	1388	18000
G3	8000	26000
G4	6002	12000
G5	2000	5987

Table 1 Dataset properties

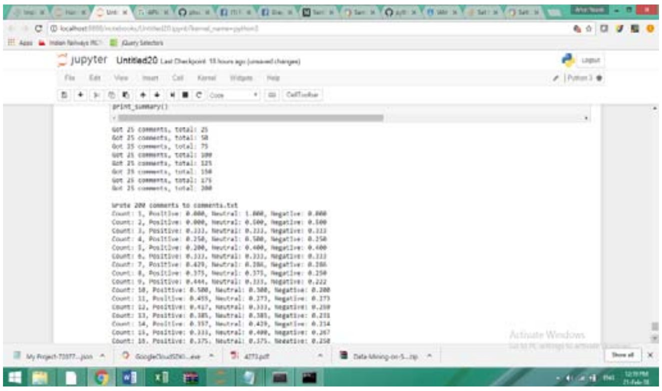
For accessing dataset from Twitter registration is necessary. After registration new project have to be created from https://app.twitter.com. Application management is used to create test app for twitter. For registration some basic information is provided like application name, organization detail, website detail. For authentication Twitter provides keys and access tokens. After getting access tokens and keys we can access the Twitter dataset.



```
Sentiment(polarity=0.0, subjectivity=0.0)
RT @oppomobileindia: RT this tweet &amp; get a chance to win the #OPPOF7 during the event! The all-new #OPPOF7 is all set to ma
ke its debut. Wi…
Sentiment(polarity=1.0, subjectivity=0.4)
RT @myselfadilkhan: @oppomobileindia Retweet done to win the #OPPOF7
JOIN in guys..
@vishalmalavya
@abizz_shaikh
@MullaFardeen
@RajPrut…
Sentiment(polarity=0.0, subjectivity=0.4)
RT @vijaytelevision: #SemiFinals இல் [illegible]
Sentiment(polarity=0.0, subjectivity=0.0)
RT @vijaytelevision: #SemiFinals இல் [illegible]
Sentiment(polarity=0.0, subjectivity=0.0)
RT @vijaytelevision: #SemiFinals இல் [illegible]
Sentiment(polarity=0.0, subjectivity=0.0)
RT @oppomobileindia: RT this tweet &amp; get a chance to win the #OPPOF7 during the event! The all-new #OPPOF7 is all set to ma
ke its debut. Wi…
Sentiment(polarity=1.0, subjectivity=0.4)
RT @oppomobileindia: RT this tweet &amp; get a chance to win the #OPPOF7 during the event! The all-new #OPPOF7 is all set to ma
ke its debut. Wi…
Sentiment(polarity=1.0, subjectivity=0.4)
RT @vijaytelevision: #SemiFinals இல் [illegible]
```

Figure 1: Semantic polarity from OPPO India Tweet

Practically all requests are delivered to the API at graph.facebook.com.For accessing Facebook dataset registration at https://developers.facebook.com/tools/explorer is required. Then access token and permissions are required to access the Facebook posts related to subjects. Facebook Graph API, and FQL Facebook Query Language can also be used for query the Facebook posts.For accessing Facebook access token is required.



```
print_summary()

Got 25 comments, total: 25
Got 25 comments, total: 50
Got 25 comments, total: 75
Got 25 comments, total: 100
Got 25 comments, total: 125
Got 25 comments, total: 150
Got 25 comments, total: 175
Got 25 comments, total: 200

Wrote 200 comments to comments.txt
Count: 1, Positive: 0.000, Neutral: 1.000, Negative: 0.000
Count: 2, Positive: 0.000, Neutral: 0.500, Negative: 0.500
Count: 3, Positive: 0.333, Neutral: 0.333, Negative: 0.333
Count: 4, Positive: 0.250, Neutral: 0.500, Negative: 0.250
Count: 5, Positive: 0.200, Neutral: 0.400, Negative: 0.400
Count: 6, Positive: 0.333, Neutral: 0.333, Negative: 0.333
Count: 7, Positive: 0.429, Neutral: 0.286, Negative: 0.286
Count: 8, Positive: 0.375, Neutral: 0.375, Negative: 0.250
Count: 9, Positive: 0.444, Neutral: 0.333, Negative: 0.222
Count: 10, Positive: 0.500, Neutral: 0.300, Negative: 0.200
Count: 11, Positive: 0.455, Neutral: 0.273, Negative: 0.273
Count: 12, Positive: 0.417, Neutral: 0.333, Negative: 0.250
Count: 13, Positive: 0.385, Neutral: 0.385, Negative: 0.231
Count: 14, Positive: 0.357, Neutral: 0.429, Negative: 0.214
Count: 15, Positive: 0.333, Neutral: 0.400, Negative: 0.267
Count: 16, Positive: 0.375, Neutral: 0.375, Negative: 0.250
```

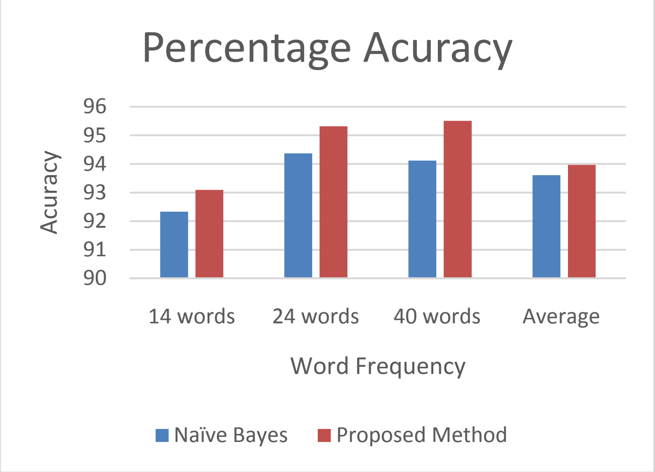
Figure 2: Facebook opinion polarity

Accuracy analysis

Attributes	% Accuracy	
	Naive Base	Proposed
10 words	92.33	93.09
20 words	94.37	95.32
30 words	94.12	95.51
Average	93.61	93.97

Table 2 Accuracy Analysis

Table above represents the accuracy of Naïve Base and proposed algorithm. As represented in table proposed method accuracy is improved as compared to Naïve Base method.



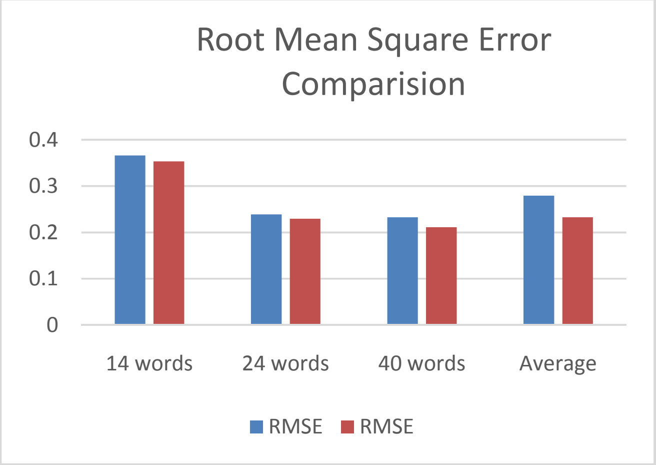
The experimental outcomes are tested with open opinions customer reviews of 450 from a twitter and Facebook users posts. The outcomes are compared percentage of precision between naIve Bayes and proposed algorithm and dissimilarity the number of feature are take out as 14,24 and 40 words respectively. The accuracy of proposed work is given values that are higher than naIve Bayes all of data sets. Moreover, the highest of accuracy value is 95.57% with 25 words and also average of proposed algorithm is higher than naIve Bayes to 94.37%.

RMSE analysis

Attributes	RMSE	
	Naive Base	Proposed
10 words	0.3660	0.3532
20 words	0.2390	0.2295
30 words	0.2326	0.2113
Average	0.2792	0.2324

Table 3 RMSE analysis.

Table 3 represented the RMSE analysis comparision of the proposed and Naïve Base method. RMSE is also reduced in our method.



The RMSE values of naIve bayes and proposed method is represented. The table above represents RMSE of different data sets. The lowest of RMSE is 40 words testing data that provide rating that are like to actual score from customer review to 0.2113. The rating of 40 words and 14 words are slightly higher value than 40 words to 0.2295 and 0.3532 respectively. The average of proposed method generates rating value that is similar actual rating as 0.2324 and median as 0.2295.

V. CONCLUSION

With the growing inspiration of online sentiment analysis and reviews on customers, the competence to detect dishonest online appraisals is crucial. The natural language processing implementations can be applied to facilitate the OM process. It provides improved natural language understanding then can help produce further improved accurate outcomes of OM. In numerous applications, it is significant to consider the context of the text data and the user preferences. The proposed method improved précised prediction for better opinion mining results. The machine learning classification technique Support Vector Machine is used for sentiment polarity. The proposed novel method can be used for semantic analysis and opinion mining to solve different sentiment polarity. The method provides automatically preprocessing of data and extract related opinion from a sentence.The data is collected from social web sites like Facebook and Twitter. The different post from users revived and mine their opinion about related subjects. The experimental outcome demonstrate that system is well suitable for accurate sentiment prediction and opinion polarity.

As a future investigation, we are concerned in considering word emotion pattern and emotion distance into text opinion computation. A microblog is posted by an individual and by only less than 135 words. It may only have a strong emotion. And a single word in dissimilar situations may have dissimilar emotion. Thus when calculating text emotion and polarity, one should thing about word context. Additionally, Artificial Intelligence computing also essentials to compute opinion or sentiment. So planning is to use transfer learning method to improve sentiment analysis.Future research may consist of an enhanced integrating of nominal meta-data. Apart from word-based content, related multimedia data can also be considered for additional study.

REFERENCES

[1] Wararat Songpan, The Analysis and Prediction of Customer Review Rating Using Opinion Mining, IEEE SERA 2017, pp. 71-77

[2] Arno Scharl, David Herring, Walter Rafelsberger, Alexander Hubmann-Haidvogel, Ruslan Kamolov, Daniel Fischl, Michael Föls, and Albert Weichselbraun, “Semantic Systems and Visual Tools to Support Environmental Communication”, IEEE SYSTEMS JOURNAL, VOL. 11, NO. 2, JUNE 2017, pp. 762-772

[3] Kamps, J., Marx, M., Mokken, R. J.*Using WordNet to Measure Semantic Orientation of Adjectives*. LREC 2004. Volume IV, pp. 1115-1118.

[4] Andreevskaia, A., Bergler, S., Urseanu, M.*All Blogs Are Not Made Equal: Exploring Genre Di_erences in Sentiment Tagging of Blogs*. International Conference on Weblogs and Social Media (ICWSM-2007), Boulder, CO. 2007.

[5] Vandana V. Chaudhari*****, Chitra A. Dhawale****** and Sanjay Misra,“ Sentiment Analysis Classification: A Brief Review”, I J C T A, 9(23) 2016, pp. 447-454

[6]ANH-DUNG VO , QUANG-PHUOC NGUYEN , AND CHEOL-YOUNG OCK, “Opinion_Aspect Relations in Cognizing Customer Feelings via Reviews”, IEEE 2017, pp. 5415-5427

[7]ATHIRA U, AND SABU M. THAMPI, “Linguistic Feature Based Filtering Mechanism for Recommending Posts in a Social Networking Group”, IEEE 2018, pp. 4469-4484

[8] S. 1. Wu, R.D. Chiang and Z.H. Ji, Development of a Chinese opinion mining system for application to Internet online forum, The Journal of Supercomputing, Springer US[Online], 2016.

[9] Z. Li, L.Liu and C.Li, Analysis of customer satisfaction from Chinese reviews using opinion mining, Proceeding of the 6th IEEE International Conference on Software Engineering and Service Science(ICSESS). 2015, pp.95-99.

[10] Q.Su, X.Xu, H.Guo, Z.Guo, X. Wu, X. Zhang and B.Swen. Hidden Sentiment association in Chinese web opinion mining. Proceeding of the 17th International Conference on World Wide Web, 2008, pp.959-968.

[11] R.M. Duwairi and I. Qarqaz, Arabic Sentiment Analysis using Supervised Classification. Proceeding of 2014 International Conference on Future Internet of Things and Cloud. 2014, pp. 579-583.

[12] H.S. Le, T.V. Le and T.V. Pham, Aspect Analysis for Opinion Mining of Vietnamese Text. Proceeding of International Conference on Advance Computing and Application, 2015, pp.118-123.

[13] V.B. Raut and D.D. Londhe, "Survey on opinion mining and summarization of user review on web", International Journal of Computer Science and Information Technology, Vol. 5(2), 2014, pp. 1026-1030.

[14] Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T.H, Kim, Opinion Mining over twiiterspace: Classifying tweets programmatically using the R approach. Proceeding of the 7th International Conference on Digital Information Management, 2012, pp. 313-319.

[15] L. Lin, 1. Li, R. Zhang, W. Yu and C. Sun, Opinion mInIng and sentiment analysis in social networks: A retweeting structure-aware approach. Proceeding of the 7th International Confernece on Utility and Cloud Computing, 2014, pp.890-895.