INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 5)
Available online at: www.ijariit.com

Sentiment analysis using machine learning classifiers and SentiWordNet: A review

Bhagyashri Sudhakarrao Wankhade
bswankhade23@gmail.com
Prof. Ram Meghe Institute of Technology and Research, Amravati, Maharashtra

Sunil R. Gupta
sunilguptacse@gmail.com
Prof. Ram Meghe Institute of Technology and Research, Amravati, Maharashtra

ABSTRACT

In the current social, technological and economic context, customers make their decisions based mostly on the opinion of other consumers. On the other side, companies require quick feedback from their customers in order to adapt to their needs in real time. The effective connection between these two aspects relies on opinion mining tools, which automatically process consumers' reviews and opinions about products or services. This paper proposes the analysis and prediction rating from customer reviews who commented as open opinion using Machine Learning Classifiers and SentiWordNet.

Keywords— *Open opinion, Customer review, Opinion mining, Naive Bayes, Support Vector Machine, SentiWordNet*

1. INTRODUCTION

With the rapid growth of Web technologies, which facilitate people to contribute rather than simply receive information, a large number of review texts are generated and become available online. These user-generated opinion rich contents are credible sources of knowledge that can not only help users make better judgments but assist manufacturers of products in keeping track of customer sentiments. However, with tens and thousands of reviews being generated every day on almost everything, e.g., sellers, products, and services, at various websites, it has become increasingly difficult for an individual to manually collect and digest the reviews of his/her interest. As such, opinion mining has become an active area of research in the past few years and has produced some important results. Online reviews have been shown to be second only to word-out-mouth in a study that compares the factors influencing purchase decisions. Therefore, online reviews can be very valuable, as collectively such reviews reflect the "wisdom of crowds" and can be a good indicator of a product's future sales performance.

Opinion mining is a type of natural language processing for tracking the sentiment or thinking of the public about a particular product. Opinion mining, which is also known as sentiment analysis, involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Automated opinion mining often uses machine learning, a component of artificial intelligence. An opinion mining system is often built using software which can extracts knowledge from examples in a database and incorporating new data to improve performance over time. The process does deep parsing of the data in order to understand the grammar and sentence structure used. Opinion mining can be useful in several ways. For example, in marketing, it is useful to judge the success of an ad campaign or new product launch, to determine which versions of a product or service are popular and even to identify which demographics like or dislike particular features. For example, a review might be broadly positive about a digital camera, but be specifically negative about how heavy it is. Capability to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups, because the data is created by the customer. An important task of opinion mining is to extract customer opinions on features of an entity. For example, the sentence, "I love the GPS function of Samsung" expresses a positive opinion on the "GPS function" of the Samsung phone. "GPS function" is the feature. The comment, "The picture of this camera is best", expresses a positive opinion on the picture of the camera. "picture" is the feature.

This paper is organized the following: the related work will be shown in section II. Section III describes the proposed methodology on how to calculate rating form customer review automatically. Finally, the conclusion is explained in section IV.

2. RELATED WORK

The opinion mining has become one of the popular research areas. The challenge is in process of opinion mining or sentiment analysis that is unstructured and noisy data on the website. A part of opinion mining refers using of natural language processing (NLP) by the proposed different method of the dictionary for sentiment analysis of text as corpus, lexicon and specific language

dictionary [4], [7], [8], [16]. They tried to extract a word from sentences for removal stop word or unnecessary word automatically. In addition, various dictionaries are solved by machine learning methods [12], [13], which try to rank scoring of various dictionaries. For example, the paper in [13] used a fuzzy logic algorithm to collect the ranking of the different dictionary into a rule for classifying the opinion. Afterword segmentation process is removal stop words by dictionary checking. The group of researches in [1], [2], [6], [9], [17] 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 the whole sentence. If the word extractions have weight from a 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 [11]. The popular classifier model is naive Bayes compared with another model [5], [8], [10], which there are different sources such as social media and website. From these researches are used classifier models that are the same objective to classified opinion. Our approach is different from them, this paper uses 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 to impact the customer who posted or commented to positive and negative.

3. METHODOLOGY

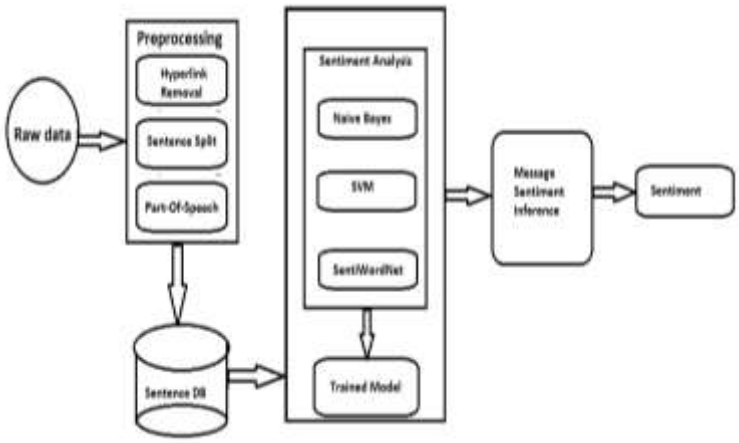

Fig. 1: Analysis and prediction of customer reviews rating using machine learning classifiers and SentiWordNet

The approach contains two major parts. Pre-processing and applying a supervised learning algorithm. In preprocessing, we remove the data to increase data consistency then we get higher accurate results. The supervised learning algorithm classifiers are Naive Bayes and Support vector machine.

3.1 Preprocessing

The pre-processing is necessary because there are some words or expressions in the review don't return any meaning and by the presence of those words, we cannot get the correct sentiment analysis. So by doing preprocessing, we get higher accurate results. In pre-processing we do the following steps.

3.1.1 Remove URLs: We remove all the links from reviews. Because they don't have any meaning. So by removing the URLs, we can get the result in minimum time. In reviews, users use these URLs to give detailed information on which he gets some idea. But for analyzing we do not require to go through all this information so we remove these URLs from our reviews.

3.1.2 Remove Repeated Letters: The repeated letters in a word are removed too. For example, we have a good word in English, we do not have words like gooooooooooood in the English language. In reviews, these words come because user like a product, then he gives the review as gooooooooooooood. In his point of view, he likes the product so much. So we remove this repeated letters and make it as good. And the word huuuuuuuungry make it as hungry because in English we have a chance that a letter can come multiple times. So we remove the letters from a word which are occurring more than twice.

3.1.3 Remove Special Symbols: We remove the symbols like : ; })] [({ etc. Because they don't have any meaning.

3.1.4 Remove Questions: We remove the questions because by getting an answer only we can analyse, so the questions are removed. For example, how r u? is a review, we remove that question because it cannot get any sentimental meaning. By Pre-processing, we get the reviews which have a complete sentimental meaning so we can easily analyse it.

3.2 Machine Learning Methods:

3.2.1 Naive Bayes: One approach for text classification is to assign to a given document d the class $c* = \arg\max_c P(c \mid d)$. We get the Naive Bayes (NB) classifier by first observing that by Bayes' rule,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

Where P(d) plays no role in selecting $c*$.
To estimate the term P(d | c), Naive Bayes decomposes it by assuming the fi 's are conditionally independent given d's class:

$$PNB(c|d) := \frac{P(c)(\prod_{i=1}^{m} P(Fi|c)^{ni(d)})}{p(d)}$$

Our training method consists of a relative-frequency estimation of P(c) and P(fi|c), using add-one smoothing. Despite its simplicity and the fact that its conditional independence assumption clearly does not take in real-world situations, Naive Bayes-based text categorization performs surprisingly well (Lewis, 1998); indeed, Domingos and Pazzani (1997) show that Naive Bayes is optimal for some certain problem classes with highly dependent features.

3.2.2 Support Vector Machine: Support vector machines (SVMs) have been shown to be most effective at traditional text categorization, but generally outperforming Naive Bayes (Joachims, 1998). They are large-margin, rather than probabilistic, classifiers, as compare to Naive Bayes. In the two-category case, the basic idea behind the training procedure is to find a hyperplane, represented by vector w~, just not only separates the document vectors in one class from those in the other, but also for which the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem; letting cj ∈ {1, −1} (corresponding to positive and negative) be the correct class of document dj, the solution can be written as

$$\underset{w}{\rightarrow} := \sum_{j} \alpha_j \; c_j \underset{d_j}{\rightarrow}, \propto_j \geq 0,$$

Where the αj's are obtained by solving a dual optimization problem. Those ~dj such that αj is greater than zero are called support vectors,s ince they are the only document vectors contributing to w→. Classification of test instances consists simply of determining which side of w→'s hyperplane they fall on. We used Joachim's (1999) SV Mlight package8 for training and testing, with all parameters set to their default values, after first length-normalizing the document vectors, as is standard (neglecting to normalize generally hurt performance slightly).

3.3 SentiWordNet Polarity

Sentiment polarity has calculated the use of SentiWordnet. SENTIWORDNET 3.0, a lexical resource explicitly devised for assisting sentiment classification and opinion mining programs. SENTIWORDNET 3.0 is a progressed version of SENTIWORDNET 1.0, a lexical aid publicly available for studies functions, now currently licensed to greater than 300studies businesses and used in a ramification of research projects international. Both SENTIWORDNET 1.0 and 3.0 are the result of routinely annotating all WORDNET synsets according to their levels of positivity, negativity, and neutrality.SENTIWORDNET 1.0 and 3.0 vary in the variations of WORDNET which they annotate (WORDNET 2.0 and 3.0, respectively), in the set of rules used for automatically annotating WORDNET, which now consists of (moreover to the preceding semi-supervised gaining knowledge of step) a random-walk step for refining the rankings [11].

3.4 Evaluation Model

The evaluation model is used k-fold cross-validation with test data which are generated all training data. The k defines the number of grouping data. For example, k is a 10-fold cross validation of 400 training data, means each group as 40 records and 10 groups, whereas the testing data will be groups 1 of 40 records and evaluation this groups to calculate the average of the accuracy collected until N as 10 groups,

$$\text{Accuracy} = \sum_{i=1}^{fold} \sum_{j=1}^{fold} \frac{\delta_{ij}}{N}$$

In addition, the results are an evaluation by rating, the root means the square error is used in this case. The comparison results are generated rating with the classifier model and rating from actual customer review as Eq.

$$\text{RMSE} = \sqrt{\frac{1}{N}} \sum_{i=1}^{n} (Pi - Oi)^2$$

where Pi is a prediction from the probability value of the classifier model. Oi is an actual score from customer review.

4. CONCLUSION

The opinion mining of customer review is very important to improve service, which the model is compared naIve Bayes and Support Vector Machine and SentiWordNet. The advantage of the classification model is calculated from the probability that is tended to the predicted class label. In additional, naIve Bayes model is able to use probability which is similar value rating, which the system is computing automatically. This paper presented a semantic approach for a sentiment analysis application, which is based on using the SentiWordNet. Even customer will be read comments, but the system can be summarized whole rating consistency with the comments. Therefore, the customers can make a decision rapidly.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] S.Atia and K. Shaalan, Increasing the accuracy of opinion mining in Arabic. Proceeding of the 1st International Conference on Arabic computing linguistics, 2015, pp.l 06-113.

[5] R.M. Duwairi and I. Qarqaz, Arabic Sentiment Analysis using Supervised Classification. Proceeding of 2014 International Conference on the Future Internet of Things and Cloud. 2014, pp. 579-583.
[6] H.S. Le, T.V. Le, and T.V. Pham, Aspect Analysis for Opinion Mining of Vietnamese Text. Proceeding of Internation Conference on Advanced Computing and Application, 2015, pp.118-123.
[7] T. Chumwatana, Using sentiment analysis technique for analyzing Thai customer satisfaction from social media. Proceeding of the 5thInternational Conference on Computing and Informatics, 2015, pp.659664.
[8] S. Ahmed and A. Danti, A Novel Approach for sentimental analysis and opinion mining based on sentiwordnet using web data. Proceeding of International Conference on Trends in Automation, Communications, and Computing Technology, 2015, pp.1-5.
[9] R.K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, Opinion mining and sentiment analysis, Proceeding of the 3rd International Conference on Computing for Sustainable Global Development, 2016, pp. 452-455.
[10] P. Barnaghim, 1.G. Breslin and P. Ghaffari, Opinion mining and sentiment polarity on Twitter and correlation between events and sentiment, Proceeding of the 2nd International Conference on Big Data Computing Service and Application, 2016, pp. 52-57.
[11] N. Kumari and S. N. Singh, Sentiment analysis on E-commerce application by using opinion mining, Proceeding of the 6th InternationalConference-Cloud System and Big Data Engineering(Confluence), 2016, pp. 320-325.
[12] V.B. Raut and D.D. Londhe, "Survey on opinion mining and summarization of user review on the web", International Journal of Computer Science and Information Technology, Vol. 5(2), 2014, pp. 1026-1030.
[13] 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.
[14] M. R. Islam and Minhaz F. Zibran, "Exploration and Exploitation of Developers' Sentimental Variations in Software Engineering", Internation Journal of Software Innovation, Vo1.4(4), 2016, pp.35-55.
[15] Y.Yokoyama, T. Hochin and H. Nomiya, "Estimation of Factor Scores of Impressions of Question and Answer Statements", Internation Journal of Software Innovation, Vol. 1(4), 2013, pp.53-66.
[16] 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 Conference on Utility and Cloud Computing, 2014, pp.890-895.
[17] A.H. Al-Hamaami and S. H. Shahrour, Development of an opinion blog mining system, Proceeding of the 4th International Conference on Advanced Computer Science Application and Technologies, 2015, pp. 74-79.

