ISSN:2320-0790

An International Journal of Advanced Computer Technology

Sentiment Analysis of Feedback Information in Hospitality Industry

Manzoor Ahmad

Scientist 'D', Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India,190001
Email: manzoor@kashmiruniversity.ac.in

Abstract: Sentiment analysis is the study of opinions, emotions of a person's towards events or entities which can enable to rate that event or entity for decision making by the prospective buyers/users. In this research paper I have tried to demonstrate the use of automatic opinion mining/sentiment analysis to rate a hotel and its service's based on the guest feedback data. We have used a semantic resource for a feature vector and Naïve Bayes classifier for the review classification after reducing the feature sets for better accuracy and efficiency. Also an improvement in the accuracy of the classification has been observed after the use of bi-gram and tri-gram language model.

Keywords: Sentiment Analysis, Opinion Mining, Naive Bayes classifiers, Language models, Features, Annotation

I. INTRODUCTION

Sentiment Analysis or opinion mining is a most researched field of study which helps to determine the quality of an event, entity or its attributes, services offered based on the users feedback or opinion. The research began with treating the problem as a text classification problem where the reviews expressed a positive, negative or neutral opinion [1]. This type of classification determined whether a sentence is subjective or objective [2]. However it becomes essential to perform a detailed analysis of a review to determine the features of a product/service that was praised or criticized by a consumer [3]. Also the consumers needed a rating for comparison with respect to the other providers for quick decision making when they want to buy a product or use a service. Sentiment analysis may be defined a computational study of collection of opinions in the form of opinionated document'd' where each opinion involves a quintuple < O,f, OO,h, t)[1,4]

1. Object (O): In general, Opinions are expressed on a target entity which can be a product, service, individual, organization, event or its attributes. Object is used to indicate the target entity on which a comment has been made.
2. Feature(f): The Object comprises which may comprises of parts or a set of attributes which are treated as features e.q room service, linen quality, cleanliness of the room. An opinion is based on one or more features of the object. The Objects are represented by a finite set of features and each feature may expressed as a finite sets of words which are synonyms of the feature.
3. Opinion Holder (h). The person who comments or expresses the opinion is the opinion holder. In case of hotel, the guest who uses a service and then expresses his opinion is the opinion holder.
4. Opinion Orientation (OO). An Opinion holder expresses his view about a feature of an object which can be neutral, negative or positive. This Positive, Negative is the orientation of that Opinion.
5. Time(t). The time at which an opinion is expressed is also important e.g the satellite channels is a feature in a hotel which may be limited at a specific time on which the guest make an opinion. But after couple of months can have a different opinion about the same feature.

Sentiment Analysis is done on a set of opinions from various users or customers. The opinions may be direct or comparative. A Direct opinion is based on a object features and includes the quintuple. The Comparative opinion expresses a preference relation of two or more object based on some of the shared features. The process of Sentiment Analysis is to evaluate these quintuples in the opinionated document and also determine the synonyms of the features in d. Lot of electronics customer feedbacks are generated every day in the form of suggestions, criticism and comments and also through elicited surveys. Taking remedial action especially when the comments are negative can impact the profitability of the units in the hospitality sector. To determine whether a feedback is negative or positive can be done by sentiment classification based on machine learning. Sentiment classification is a type of text classification where the criterion of the classification is the attitude expressed in the text rather than the content or the topic. [6] So the words which express the negative or the positive attitude needs to be taken care of. Semantic classification is done based on semantic or affect lexicon [7] or a large scale knowledge base [8] and also using machine learning from the tagged data [9].

II. NAÏVE BAYES CLASSIFIER

Naïve Bayes classifiers are probabilistic classifiers and are based on applying Bayes theorem with independent assumptions between the features. Naive Bayes involves simplifying conditional independence assumption for text classification where in the problem of deciding whether a document can belong to one category or the other with word frequencies as the features. Maximum Likelihood naive Bayes classifiers can be trained very efficiently in a supervised learning setting. An optimal classifier can be built if we know the priors $P(\omega_i)$ and the class-conditional densities $P(x|\omega_i)$ and have some knowledge and training data$\{(x_i, \omega_i)\}$and use the sample to estimate the unknown probability distributions. Also x is typically high dimension and we have to estimate$P(x|\omega)$) from limited data. The maximum likelihood probability of a word belong to a particular class is given by the expression. Given the categories $\{\omega_1, \omega_2, \omega_3, \dots \dots \omega_c\}$, the feature vector $X = \{x_1, x_2, x_3, \dots \dots x_d\}^t$, the Naïve Bayes assumption is given by

$$P(x|\omega_i) = \frac{count\ of\ x_i\ in\ documents\ of\ class\ \omega_i}{Total\ number\ of\ words\ in\ documents\ of\ class\ \omega_i}$$

Given the categories $\{\omega_1, \omega_2, \omega_3, \dots \dots \omega_c\}$, the feature vector $X = \{x_1, x_2, x_3, \dots \dots x_d\}^t$, the Naïve Bayes assumption is given by

$$P(x_1, x_2, x_3, \dots \dots x_d|\omega_j) = \prod_i P(x_i|\omega_j)$$

And the Naïve Bayes Classifier is given by

$$\omega_{NB} = arg\ _{max}\ P(\omega_j) \prod_i P(x_i|\omega_j)$$

The model uses simplifying conditional independence which means that the words are conditionally independent irrespective of the class. The classifier outputs the class with maximum posteriori probability. In case the classifier comes across a word out of the training set the probability of both the classes become zero. Lapace smoothing is used to overcome this issue which is given by

$$P(x|\omega_i) = \frac{count(x_i) + k}{(k+1) * Number\ of\ words\ in\ class\ \omega_i}$$

The k is chosen to be equal to 1 .

III. DATA

The sentiment analysis is conducted on the review type data consisting of small paragraphs of text as well as survey type of data. For the classification we have used semantic resource for the opinion classification as positive, negative or neutral and then used machine learning technique on the guest feedback to analyze the distribution of the lexical elements for correct conclusions. A semi-supervised technique has been followed wherein a Naïve Bayes Classifier has been trained on a set of pre-tagged data and then the performance of the classifier was evaluated on the test data. The data set comprised of the 14554 reviews collected on the web for a hotel with the indicator levels for different features as well as offline feedbacks in the electronic form. A feature vectors comprising of a different combination of features were selected for performance evaluation as well to determine the correct set of features and the improvement in the accuracy of the sentiment analysis.

IV. CLASSIFICATION OF GUEST FEEDBACK

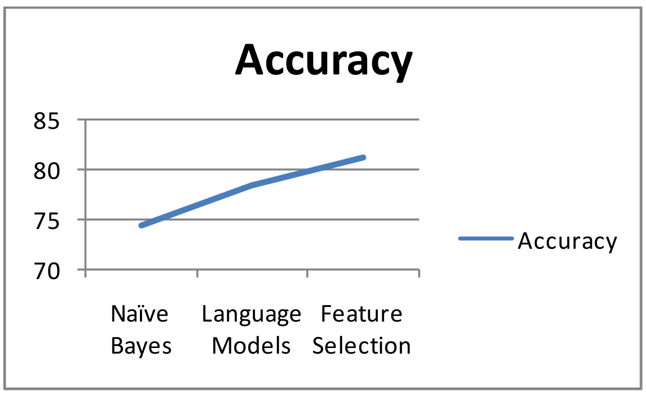
The Opinion or the feedback is conveyed by adjectives or a combination of adjectives with other parts of speech. Using language models we have tried to overcome this problem to identify word like “The services are very good” or The hotel is definitely recommended for a business class traveler”. By using a bi-gram or tri-gram language model we captured this information and increased the accuracy of the analysis for determining the whether a feedback is negative or positively biased. Feature selection has been performed to remove the redundant features and at the same maintaining those features which have very high disambiguation capabilities. Feature reduction can optimize the performance of a classifier by reducing the feature vector to a size that does not exceed the number of training cases. Feature reduction is done either by elimination of set of features based on linguistic analysis or by selecting the top ranking ‘n’ features based on some criterion of predictiveness [6]. We used a hybrid approach and the measure of predictiveness was log likelihood with respect to the target variable [7] We experimented with different feature sets so as to establish a gain in the sentiment classification task by selecting a qualitative set of features which contributes to the same in terms of efficiency as well as accuracy.

V. RESULTS AND PERFORMANCE

We implemented the classifier in java wherein we used a training data set for classifier and then tested on the test data set of a hotel. The optimal numbers of features were selected by testing the classifier on the test data. We obtained an overall accuracy of 81.2% on the test data. The results may not be as satisfactory as maximum entropy classification or support vector machine but the algorithm can be modified to use the linguistic analysis for increasing the accuracy.

Experiment	Results
Naïve Bayes Classification	74.3%
Naïve Bayes Classification with Language models	78.4%
Naïve Bayes Classification with feature selection	81.2%

Table 1 : Results



VI. CONCLUSION

We have performed the sentiment analysis on the hotel feedback using Naïve Bayes Classifier which is normally thought to be less accurate. But we have shown that on this hotel feedback data the results can be compared to the classification accuracy of the more advanced machine learning algorithms. Naïve Bayes Classifier being a probabilistic model is fast while training can be used for large data sets and less prone to over fitting. By using an bi-gram and tri-gram language model the accuracy of the classifier increased. Also using initially large feature set combined with feature reduction also contributed to the classification accuracy in the sentiment analysis. The ideas in this paper can be extended for other types of data sets and used with sophisticated machine learning algorithms.

VII. REFERENCES

[1] B.Liu, "Sentiment Analysis and Subjectivity, Handbok of Natural Language Processesing". Second Edition(editors: N. Indurkhya and F. J. Damerau), 2010

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis ". Foundation and Trends in Information Retrieval 2(1-2), pp. 1-35, 2008.

[3] J. Wiebe, T. Wilson, R. Bruce and M. Martin, "Learning Subjective Language", Computational Linguistics, Vol 30, pp 277-308, September 2004.

[4] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews", Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp, 168-177,2004.

[5] Michael Gamon, " Sentiment classification on customer feedback data : noisy data, large feature vectors and the role of linguistic analysis", Proceeding of COLING-04, the 20th International Conference on Computational Linguistics, pp 841-847 , Geneva, CH

[6] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack,"Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques", Proceeding of ICDM-03, the 3ird IEEE International Conference on Data Mining pp 427-434, Melbourne, US

[7] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[8] Dave, K., Lawrence, S., and Pennock, D.M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. WWW 2003

[9] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.

[10] Rennie, Jason D., et al. "Tackling the poor assumptions of naive bayes text classifi-ers."Machine learning international workshop and conference Vol. 20. No. 2. 2003

[11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, DanHuang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th An-nual Meeting of the Association for Computational Linguistics (ACL 2011).

[12] Kennedy, Alistair, and Diana Inkpen. "Sentiment classification of movie reviews using contextual valence shifters."Computational Intelligence22.2 (2006): 110-125.

[13] Aidan Finn and Nicholas Kushmerick (2003): „Learning to classify documents according to genre". IJCAI-03 Workshop on Computational Approaches to Text Style and Synthesis.

[14] Li, Tao, Yi Zhang, and Vikas Sindhwani. "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge."Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. Associationfor Computational Linguistics, 2009.

[15] Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In PAKDD 2005 (pp. 301–311).Springer-Verlag: Berlin, Heidelberg.

[16] Whitelaw, Casey, Navendu Garg, and Shlomo Argamon."Using appraisal groups for sentiment analysis." Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.

[17] Peter D. Turney (2002): "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In: Proceedings of ACL 2002, pp. 417-424