ISSN:0975-9646

An Approach to Perform Aspect level Sentiment Analysis on Customer Reviews using Sentiscore Algorithm and Priority Based Classification

Aishwarya Mohan, Manisha.R, Vijayaa.B, Naren.J

Computer Science Department, School of Computing, SASTRA University
Thirumalaisamudiram, Thanjavur, India.

***Abstract*— This paper analyses the customer reviews on restaurant domain using sentiment analysis and text mining techniques. The most integral part of our work is to assign Sentiment scores to the aspects with respect to the words used. We have devised Sentiscore algorithm to perform this function. The dataset we have at our disposal is a set of review documents obtained from an authenticated repository. We perform an aspect level sentiment extraction thereby, attempting to mine and understand the user's feedback data. The aspects that we have taken into account are food, cost, ambience and service. A priority-based algorithm forms the rule base for the classifier to predict the polarity of the reviews. To start with, we perform a clean up on the review data. Sentiscore algorithm is then utilized to generate the aspect document matrix. Polarity prediction is performed using Naïve Bayes and k-nn classifier and the results are evaluated. Finally, the experimental analysis shows that, k-nn performs better with increasing number of instances. The proposed methodology will equip the restaurant owner to identify the areas that require improvement.**

***Keywords*— Text Mining, Aspect-level Sentiment Analysis, polarity prediction.**

I. INTRODUCTION

Customer reviews have become an important source of information nowadays, with companies trying to understand customer provided feedback. User review is essential in almost all the fields. With more and more common users becoming comfortable with the Web, an increasing number of people are writing reviews on it. As a result, the number of reviews a system receives grows rapidly. Further, the internet slang words make it difficult to comprehend .The present day review analysis systems mostly determine the polarity of a sentence as a whole. However the reader might be interested to know the aspect-wise summary of opinions rather than the overall picture.

Sentiment analysis is the process of analyzing text to identify positive and negative opinions. This paper presents the analysis of unstructured data using restaurant reviews as a case study. Generally the restaurant owner would like to know how well his system functions. This is accomplished by performing sentiment analysis on the review data. Our proposed system gives him an aspect wise summary rather than the overall sentiment of a review.

II. RELATED WORK

Sentiment analysis has been an area under study in the recent years. Many researchers have contributed a lot in the development of this study. Bing Liu [1] is one among the notable contributors. His book on Natural Language Processing has focused on the mining of huge volume of texts with opinions. Another researcher Li Zhang, Weiran Xu, Si Li [2], proposed a method for object extraction and aspect identification using a slack function. He used dictionary based method to analyze the sentiment for the aspect. Su Su Htay and Khin Thidar Lynn [3], proposed a novel idea making use of Pattern Knowledge to find the sentiment words. Customer reviews are taken as input and opinion summary is produced as output in this paper. Theresa Wilson, Janyce Wiebe and Paul Hoffmann [4], together designed a system to automatically identify the contextual polarity for a huge set of sentiment terms, achieving phrase-level sentiment analysis. An algorithm on opinion orientation was presented by Xiaowen Ding, Philip S. Yu [5] which was used to provide a holistic lexicon-oriented methodology to solve the problem by utilizing linguistic rules and external evidences. Chee Kian Leong a, Yew Haur Lee b, Wai Keong Mak [6] devised a system that is capable of mining SMS texts with emoticons. S.L. Ting, W.H. Ip, Albert H.C. Tsang [7] highlighted the performance of naïve bayes classification in their paper, thus showing that it is one of the simplest classifiers to learn and implement. M. Ikonomaki, S. Kotsiantis, V. Tampakas [8], in their paper, illustrated the various machine learning algorithms that can be used to perform classification. Overall flow is given below.

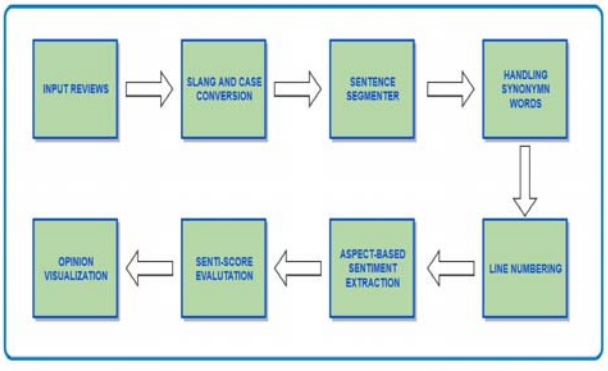

Fig. 1 Overall flow diagram

III. METHODOLOGY

Having considered Restaurant review domain, we are concentrating on predefined aspects such as **food, ambience, cost and service.** We let the user (restaurant owner in this case) to give his/her priority on the aspects. This is done to get a holistic picture of the classification of reviews based on one's preference and priority.

A. Text-Pre-Processing

Pre-Processing of text involves cleaning the data and retrieving essential information for further processing. Its basic task is to prepare the text as per the needs for the forth-coming steps. Some of the common pre-processing techniques generally employed are Stop-Word removal, slang and case conversion, stemming etc.

The dataset obtained from the repository is a collection of restaurant reviews. Each review consists of several sentences. Reviews these days have lots of slang words like "Gr8","Awsum". Thus we intend to perform **slang conversion and case conversion** in the very beginning. Each review is broken down into sentences using a **sentence segmenter** (based on delimiters). Synonymous words for the considered aspects are handled. Rules are defined to split sentences based on conjunctions in order to separate the multiple aspect sentences. The sentences are now numbered from 1.1, 1.2...2.1, 2.2...and so on till the final sentence in the document. The first number indicates the review number and the second number indicates the sentence number within a particular review. **Aspect based sentence extraction** is done by looking for sentences having the aspects (food, service, cost, ambience) in the file and putting them into their respective text files like "food.txt","service.txt","cost.txt" and "ambience.txt".

B. Score-board Generation

Scores are to be assigned next. We have devised Sentiscore algorithm to associate scores to aspects based on the intensity of the sentiment words used. The algorithm makes use of four word-lists: idiom list, negative word list, booster list and emotion-word list. (Words are taken from Harvard inquirer lexicon and revised for restaurant domain with a score range of -4 to +4). The algorithm extracts the sentences one by one to look up against the four word lists (in the same order) and assign scores appropriately. Idioms if present are assigned scores at the start. The polarity of an emotion word is reversed in case a negative word is present. For Example: "not good" is assigned a negative polarity. The presence of booster words like "very", "much"…intensifies the score of an emotion word by +1 or -1 based on its polarity. Comparative sentences like "hospitality quotient was better than the food" are also handled taking the comparative word into consideration. For every iteration, the polarity score variables (pos and neg) are updated based on the highest positive and negative scores encountered. Finally the results are written to a file in the format mentioned in the algorithm.

1) *Sentiscore Algorithm:*

```
FOR every File
  WHILE(EOF)
    FOR every sentence s,
        set pos to 0
        set neg to 0
        Check if idiom is present in s,
        l=s
        IF yes,
            l=s without idiom
            update pos , neg based on idiom
        END IF
        Tokenise l,
        FOR every token t
          Check if t is a neg-word,
          IF yes,
              Check if next word is an emotion-word.
              IF yes,
                 Extract score.
                 invert the scores.
                 Update pos and neg based on the magnitude of scores.
              END IF
          END IF
          Check if t is a booster word,
          IF yes,
              Check if next word is an emotion-word.
              IF yes,
                 Extract score.
                 if positive -> add 1 to emotion-word score.
                 if negative-> sub 1 from emotion-word score.
                 update pos and neg based on the magnitude of scores.
              END IF
          END IF
          Check if t is neither a booster-word nor neg-word.
          Check if t is an emotion-word.
          IF yes,
              Extract score
              Update pos and neg based on the magnitude of scores.
           END IF
        END FOR
        IF either pos or neg is non-zero
          Write line into the output file in this form
          Pos<TAB_SPACE>neg<TAB_SPACE>s
        END IF
    END FOR
  END WHILE
END FOR
```

An aspect document matrix is constructed using the scores obtained. Each row of the matrix represents a review and columns represent the aspects. The individual aspect scores for each document is used for determining the final class label value (Positive or negative).

C. *Classification Process*

Classification in data mining is used to predict class label for data instances. Aspect Document Matrix is subjected to classification using two common classifiers: Naive Bayes and k-nn. Comparison of the performance between the two classifiers is done. A **Naive Bayes classifier** is a probabilistic classifier making use of Bayes theorem with strong independence assumptions. Naive Bayes classifiers are simple and can be trained easily in supervised learning. The classifier's striking feature lies in its computational efficiency, simplicity and good performance. **K-nn or lazy learning** is an instance-based learning, where the function is approximated with respect to neighbourhood and all calculations are postponed until classification. We have defined rules which make use of the aspect priority input from the user to predict the class label. For example, a user wishes to have food as top priority followed by service, cost and ambience. If the review contains positive emotion towards the top-priority aspects (food and service) and strongly negative emotions towards cost and ambience, the class label will be predicted as positive based on the rules defined. A confusion matrix depicts the classifier performance results.

D. *Results and Discussion*

Table I shows the accuracy, precision, recall and FMeasure values for the classifiers based on increasing number of instances.

TABLE I CLASSIFIER RESULTS

CORPUS TAKEN	CLASSIFICATION MODEL	ACCURACY	PRECISION	RECALL	FMEASURE
N=20	NB	0.80	0.80	0.80	0.79
	KNN	0.75	0.744	0.75	0.745
N =40	NB	0.825	0.829	0.825	0.826
	KNN	0.85	0.85	0.85	0.85
N=200	NB	0.85	0.851	0.85	0.85
	KNN	0.865	0.866	0.865	0.865
N=500	NB	0.934	0.934	0.934	0.934
	KNN	0.944	0.944	0.944	0.944
N=1000	NB	0.946	0.947	0.947	0.947
	KNN	0.966	0.966	0.966	0.966

Based on the priority rules defined t he Naïve-Bayes classifier trains itself to perform better when the number of instances are small. However, as the number of instances increases its performance drops in comparison to k-NN.

Fig.2 clearly shows that k-nn trains itself to perform well on a large dataset. And, to get a holistic picture of a system large number of reviews are necessary. Fig.3 shows that the feedback of the restaurant system taken into consideration is predominantly positive. Fig.4 highlights the aspects which require further enhancement. Hence the restaurant owner can work on the system accordingly, thereby raising its standard.

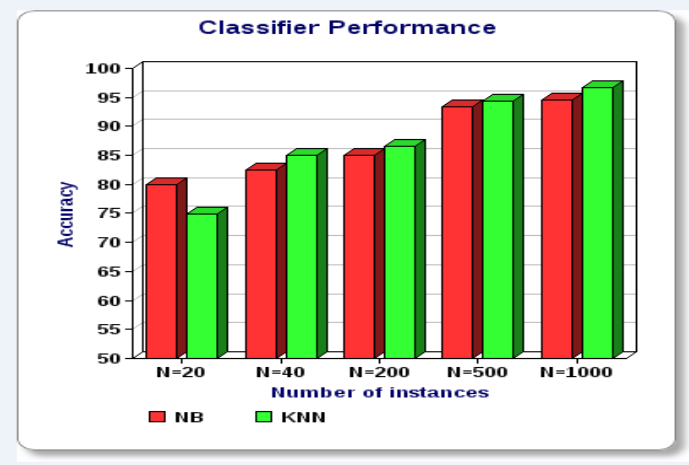

Fig.2. Performance Visualisation

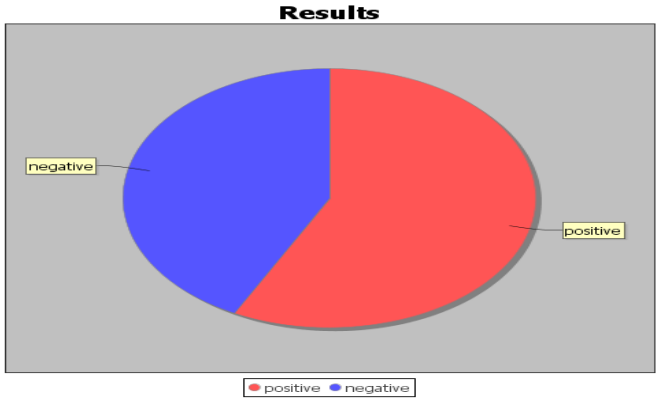

Fig.3. Overall rating based on priority classification

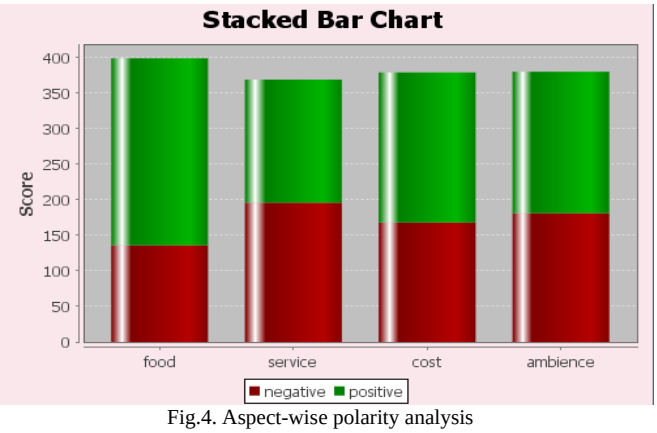

Fig.4. Aspect-wise polarity analysis

IV. CONCLUSION AND FUTURE WORK

In this paper, we have applied the Text Mining and Sentiment Analysis techniques on reviews which can be expanded to any domain. These techniques are employed to analyze the feedback given by people for improving the performance of the system. Existing Systems are generally capable of obtaining the polarity for sentences as a whole. The proposed method is a novel one that determines the polarity of every aspect in a multi-aspect sentence. Our work provides guidelines for the restaurant owner highlighting the areas which require improvement based on the reviews collected.

A potential area of future research is handling of sarcasm within reviews. Another improvement would be to take timestamp information into consideration, so as to analyze the changes in opinions about aspects over a period of time. Aspect-extraction can also be done directly without any need to predefine them explicitly.

REFERENCES

[1] Liu. B, *Sentiment analysis and subjectivity*, In N. Indurkhya & F. J. Damerau (Eds.), Handbook of natural processing language (pp. 627-666), Boca Raton: CRC Press, 2010.

[2] Li Zhang, Weirman Xu, Si Li, *Aspect Identification and sentiment analysis based on NLP*, Proceedings of IC-NIDC 2012, 660-664, 2012.

[3] Su Su Htay and Khin Thidar Lynn, *Extracting Product Features and Opinion Words using Pattern Knowledge in customer reviews*, Hindawi Publishing Corporation The scientific World Journal, Article ID 394758, (1-5), 2013.

[4] Theresa Wilson, Janyce Wiebe and Paul Hoffman, *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.*

[5] Xiaowen Ding, Philip S. Yu , *A Holistic Lexicon Based Approach to Opinion Mining*, 2007.

[6] Chee Kian Leong, Yew Haur Lee b, Wai Keong Mak, Mining sentiments of SMS texts for teaching evaluation, Expert Systems with applications 39:2584-2589, 2012.

[7] S.L. Ting, W.H. Ip, Albert H.C. Tsang, Is Naive Bayes a Good Classifier for Document Classification? International Journal of Software Engineering and Its Applications Vol. 5, No. 3, 2011.

[8] M. Ikonomaki, S. Kotsiantis , V. Tampakas, *Text Classification Using Machine Learning Techniques*, WSEAS transactions on computers, Issue 8, Volume 4, , pp. 966-974, 2005.

[9] Dave, K., Lawrence, S., and Pennock, D, *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW'03*. ., 2003

[10] Tsung-Hsien Chiang, Hung-Yi Lo, Shou-De Lin, *A Ranking-based KNN Approach for Multi-Label Classification*, 2012.