Opinion Mining Classification Using Key Word Summarization Based on Singular Value Decomposition

B Valarmathi*
Department of Computer Science and Engineering
SKP Engineering College
(Affiliated to Anna University, Chennai)
Tiruvannamalai – 606 611
* corresponding author

Dr.V Palanisamy
Principal
Info Institute of Engineering
(Affiliated to Anna University, Coimbatore)
Coimbatore – 641 107

Abstract—With the popularity of online shopping it is increasingly becoming important for manufacturers and service providers to ask customers to review their product and associated service. Typically the number of customer reviews that a product receives grows rapidly and can be in hundreds or even thousands. This makes it difficult for a potential customer to decide whether to buy the product or not. It is also difficult for the manufacturer of the product to keep track and manage customer opinions. Opinion mining is an emerging field that classifies a user opinion into positive and negative reviews. In this paper it is proposed to develop a methodology using word score based on Singular Value Decomposition by modeling a custom corpus for a given topic in which opinion mining has to be performed. Bayes Net and decision tree induction algorithms are used to classify the opinions.

Keywords; Opinion mining, classification, Bayes Net,CART, Sequential minimal optimization, movie review.

I. INTRODUCTION

An emerging area of research is the modeling of opinion, sentiment, and subjectivity which has recently attracted a great deal of attention because of its potential applications [1]. Text data in general can be broadly classified into facts and opinions. Facts are objective statements whereas opinions are subjective statements that reflect on a person's perception about an event or entity [2]. Most of the existing research on text information processing, web information processing focused on information retrieval in the factual domain rather than the opinion domain.

Opinion mining is a recent discipline which extensively uses information retrieval and computational linguistics and is not concerned with the topic of a text, but the opinion it expresses. Opinion mining has applications ranging from determining critics' opinions about a given product by classifying online product reviews, or tracking the shifting attitudes of the general public towards a movie star by mining online forums or blogs[3][4].

Opinion classification has been widely studied by the natural language processing community [5,6] and is defined as follows: Given a set of text data D, it analyzes whether each document $d \in D$ expresses a positive or negative opinion on a specific object. For example, given a set of blogs on movie reviews, the system classifies them into positive reviews and negative reviews. This is similar to a supervised classification method but different from the regular topic based text classification, which classifies documents into predefined topic classes, e.g., sports, art etc. In topic-based classification, topic related words are important. However, in opinion classification, topic-related words are not very important but, opinion words that indicate positive or negative opinions are important, e.g., great, excellent, amazing, horrible, bad, worst, etc. Most of the methodologies for opinion mining apply some forms of machine learning techniques for classification. Customized-algorithms specifically for opinion classification have also been developed, which exploit opinion words and phrases together with some scoring functions [7].

In this paper we investigate classification of opinion mining not only based on opinion words but also corpus words which are frequently used in the documents under review. We also propose a methodology to eliminate key words that are commonly used in the dataset under study. For example the word "movie" is irrelevant for classification of movie reviews. We rank the corpus using Singular value decomposition and prepare our data for opinion mining.

This paper is organized into the following sections. Section II describes the dataset used in our work, section III briefly describes the classification algorithms, section IV describes the experimental setup and section V analyzes the obtained results with conclusion.

II. DATASET USED

It was proposed to work with movie reviews due to the availability of a large number of reviews available online. Bo Pang and Lillian Lee [8] provide collections of movie-review documents labeled with respect to their overall sentiment polarity (positive or negative). This was selected because Turney[7] found movie reviews to be the most difficult of several domains for sentiment classification and state "It appears that movie reviews are difficult to classify, because the whole is not necessarily the sum of the parts; thus the accuracy on movie reviews is about 66%". We use 150 positive and 150 negative opinions from the polarity dataset version 2. Our work stresses on the methodology rather than a specific domain and hence can be used for any other review dataset also.

The data source used by Bo Pang and Lillian Lee was the Internet Movie Database (IMDb) archive of the rec.arts.movies.reviews newsgroup. Reviews were selected only where the author rating was expressed either with stars or some numerical value. Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral. However in this paper we focus only on the positive and negative reviews.

III. CLASSIFICATION ALGORITHMS

Classification is a supervised technique with labeled examples for the class attribute which is used as the training set by the classification algorithm and the unlabeled example for the class attribute which needs to be found using the multiple predictor attributes available. Classification accuracy depends on the model being built using the historical data that accurately predicts the label (class) of the unlabeled examples. Popular techniques include Bayesian approach, Decision tree induction approach, Support vector machine and Neural network approach.

A Bayesian network is a probabilistic graphical model that describes a group of random variables and their conditional dependencies (defined using conditional probability table) via a directed acyclic graph (DAG). In a Bayes Net classifier, edges in the DAG represent conditional dependencies whereas nodes which are not connected represent attributes which are conditionally independent.

For a Directed acrylic graph, *DAG* = (*V*,*E*) , if $Y = (Y_x)$ where $x \in X$ be a set of random variables indexed by X. Y is a Bayesian network with respect to the *DAG* when the joint probability density function can be calculated as the product of the individual density functions and conditional on their parent variables and is given by

$$p(y) = \prod p(y_x \mid y_{pa(x)})$$

pa(x) is the set of parents of x.

Classification and regression trees is a recursive partitioning method used to predict continuous dependent variable and categorical variables based on the target variable. CART follows the typical decision tree induction method, where a major challenge is to identify the variable split criterion which
has a major impact on the quality of the resulting tree. The goal of splitting up a sample is to get sub-samples that are more pure than the original sample. A commonly used technique is to choose a split that will create the largest and purest child nodes by only looking at the instances in that node. In this paper we use the Gini impurity criteria for splitting the node given by

$$l(t) = \sum_{i=j} p(i \mid t)\, p(j \mid t)$$

P(j|t) is the conditional probability of having j class in t node.

IV. EXPERIMENTAL SETUP

Of the 150 positive and 150 negative reviews selected, the first step involves selecting all words found in the input documents. This involves indexing and counting to compute a table of documents and words, i.e., a matrix of frequencies that enumerates the number of times that each word occurs in each document. This basic process was further refined to exclude certain common words such as "the" and "a" (stop word lists) and to combine different grammatical forms of the same words such as "traveling," "traveled," "travel," etc.

Since the above method also retrieves common words used in a specific type of dataset, a corpus was created which becomes the include list for the specific type of dataset. The corpus creation is described by the rules given below.

1. If $w_i \in W$ and $\sum w_i f_i \geq 75\%$ of initial word population, update exclude list.
2. If $w_i \in W$ and $\sum w_i f_i \leq 15\%$ of initial word population, update exclude list.

Based on the above rule the word frequency is created using the exclude list. Singular value decomposition is used to find the importance of the word. The purpose of SVD is to reduce the overall dimensionality of the input matrix to a lower dimension space where each consecutive dimension represents the largest degree of variability between the selected word and documents.

Scoring is done on the words to create the test data by factoring the word importance based on SVD and the word frequency. The numerical data so obtained from the above process for each word is used to train the classification algorithm. Figure 1 and Figure II shows the most and least common word used for our analysis.

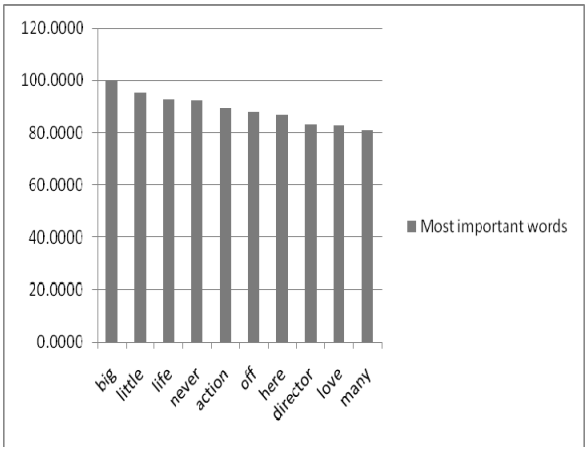

Figure I : Most important words extracted

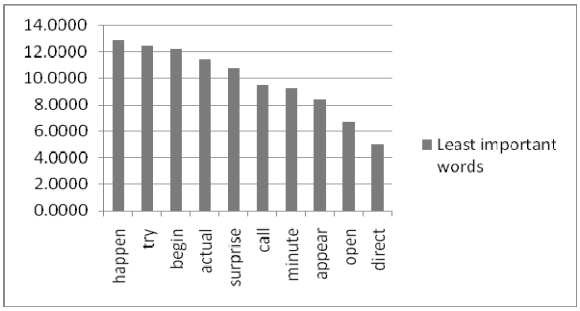

Figure II : Least important words extracted

From figure 1 we observe that words including “off”, “action” plays a important role for classifying opinion whereas words like “actual” does not play a crucial part in the classification.

The data prepared in the above method was preprocessed using sequential minimal optimization and is used to train the Bayes Net classification algorithm and Classification and regression tree algorithm. A 10 fold cross validation was used during the test mode. The classification results obtained from both the methods is shown in figure III.

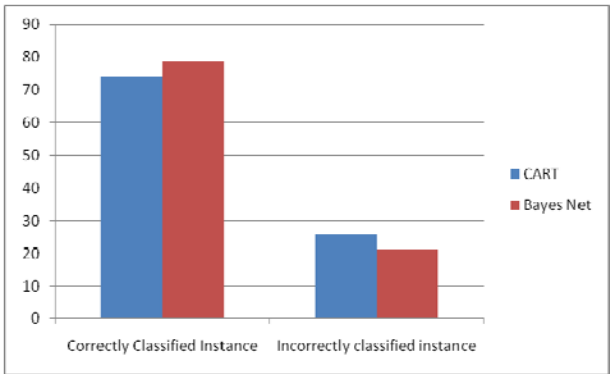

Figure III. Classification accuracy

The true positive and false positive for the respective algorithms are shown in figure IV

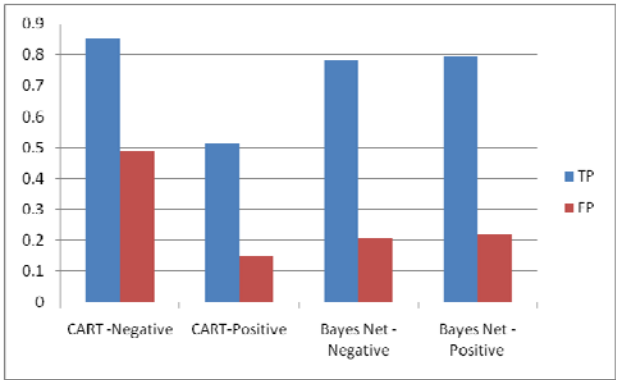

Figure IV. TP and FP rates

V. CONCLUSION

In this paper we analyze a novel method of creating exclude list from the extracted words in the document. The corpus of words created after the exclude list was created and given scores based on SVD. CART and Bayes Net with 10 fold cross validation was used to determine the classification accuracy. The output obtained was 76% and 78.667% respectively. The proposed method shows pretty good results. Work needs to be carried out on other datasets to check the proposed methods accuracy on different types of datasets.

REFERENCES

[1] Beineke, Philip, Trevor Hastie, Christopher Manning, and Shivakumar Vaithyanathan. 2004. Exploring sentiment summarization. In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI tech report SS-04-07).

[2] Bing Liu. Opinion Mining.

[3] Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, pages 617–624, Bremen,DE.

[4] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, pages 271–278,

[5] Dave, D., Lawrence, A., and Pennock, D. Mining the Peanut Gallery: Opinion Extraction andSemantic Classification of Product Reviews. Proceedings of International World Wide Web Conference (WWW'03), 2003

[6] Pang, B., Lee, L. and Vaithyanathan, S. Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02), 2002.

[7] Turney, P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL'02, 2002.

[8] Movie Review Data "http://www.cs.cornell.edu/People/pabo/movie-review-data/"