ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

SO-PMI Based Sentiment Analysis with Hybrid SVM Approach

Vinay Shivaji Kamble, Sachin N. Deshmukh

M.Tech Student, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

Professor, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

ABSTRACT: The huge amount of unstructured content generated by user has important part in rapid growth of research in text mining for sentiment analysis. This paper proposes a different model for sentiment analysis of movie reviews using a combination of PMI based algorithm and other algorithm. Firstly, different text data pre-processing schemes are applied on the dataset. Secondly, the different classifiers such as SO-PMI, Word Based Approach is applied and also SVM‘s behaviour is analysed in combination with PMI based algorithm to obtain the different results for sentiment analysis to increase the result accuracy.

KEYWORDS: Opinion mining, Pointwise Mutual Information, Semantic Orientation, Sentiment Analysis

I. INTRODUCTION

Sentiment analysis is the process of determining the contextual polarity of the text i.e. whether a text is positive, negative or neutral. Extraction, identification or otherwise characterization of the sentiment content in the text unit using statistics and methods of machine learning are also referred as Sentiment Analysis or Text Analysis [1]. The area of sentiment analysis and opinion mining has seen a large growth or increase in scholar's interest in the past few years. Scholar's in the areas of NLP i.e. natural language processing, data mining, machine learning, and others have tested a various methods of automating the sentiment analysis process [2]. Text classification is the process of assigning certain categories to text documents or text documents are classified on the basis of certain categories. The task of classifier is to define the appropriate category or class for each text document based on the input model or algorithm used for classification. Emerging new trends in the field of internet and computers increases the processing of text data at a given time, due to this, there is a need for maintaining and organizing these data to provide easy storage and access. Many text classification approaches with increased accuracy were developed for effectively identification and classification problem of these data [3]. In this proposed work, new hybrid classification method is proposed based on joining or combining classification methods and their performances result are analysed in terms of accuracy measurement. The sentiment or opinion found within user comments, their feedback or critiques provides helpful information for many different purposes. These sentiments can be classified either into two categories: Positive and Negative; or into an N-point scale, e.g., very bad, bad, satisfactory, good, very good. In this respect, a sentiment analysis or opinion mining task can be interpreted as a classification task where each category represents opinion or sentiment. Sentiment analysis provides business-holder with a means to calculate the extent of product acceptance and to determine new strategies to improve product quality and efficiency. It also make easier for policy makers or politicians to determine public sentiments views with respect to different policies, public services or political issues [4].

Sentiment Analysis and Opinion mining for text document have recently made a lot of attention because of their many useful applications for business-holders and organizations such as extracting of customer's sentiment, fully automated recommender systems, or deducing public sentiment or opinion about a certain topic. Two main goals of opinion or sentiment mining are: (i) to evaluate whether a given text contains any opinion as opposed to being factual or objective, and (ii) for extracting the sentiment of a given text by classifying it as positive, negative, along with neutral with respect to the given target . Furthermore, opinion mining can be performed at the sentence or document level. In general, sentence-level mining is made very complex by the fact that the semantic orientation of words is highly context-dependent, and document-level mining is made very complex by the fact that one document may contain several contradictory opinions or sentiments about the same target [5].

II. RELATED WORK

The researches or scholar in the field of Sentiment Analysis started much earlier in 1990‟s but in the year 2003, the terms Opinion Mining and Sentiment Analysis were first introduced. The earlier work is focused on subjectivity detection, interpretation of metaphors and sentiment adjectives, so this was very limited [6].

In experiments with 410 different reviews from Epinions, the algorithm gets an average accuracy of 74%. It observes that movie reviews are difficult for classification, because the whole review having many different factors that also have to consider; therefore the accuracy on movie reviews is about 66%. On the other hand, for the reviews of banks and automobiles, it appears that the whole is the addition of all the parts, and the accuracy is ranging from 80% to 84%. Reviews about Travel are an intermediate case. For predicting semantic orientation Hatzivassiloglou and McKeown in 1997 have also developed a new algorithm. Their algorithm performance is good, but it was developed for an isolated adjectives, rather than phrases containing adjectives or adverbs. This algorithm classifies adjectives with accuracies ranging from 78% to 92%, depending on the amount of training data that is available [7].

III. DATA COLLECTION, PREPROCESSING & METHODOLOGY

A. Data Collection and Preprocessing

Data set [8] consist of 1009 text document of movie review. In data pre-processing, all text was converted to lowercase, for making the data same. Then meaningless words are converted to meaningful words wherever possible. e.g. goood converted to good, unwanted punctuations such as comma, numbers i.e. un-necessary data should be remove. After this we have to perform POS tagging.POS tagging is nothing but the part of speech tagging in which we tag each word to get its part of speech. It will help us to select particular word of particular part of speech.

After pre-processing work, feature extraction process is carried out. In feature extraction we have to extract the phrases of particular pattern appear in the sentence/text document .The pattern is as follows which is used by Turney in 2002[7].We used POS tagging method as earlier discussed.

Table-1: Feature selection pattern

Word1	Word2
JJ	NN or NNS
RB,RBR or RBS	JJ
JJ	JJ
NN or NNS	JJ
RB,RBR or RBS	VB,VBD,VBN or VBG

B. Methodology for Sentiment Analysis

1. SO-PMI based approach:

This method calculates the PMI i.e. Point-wise Mutual Information between two words and produce numeric score. The formula is as follows.

$$\text{PMI (word1, word2)} = log_2\left(\frac{\text{prob(word1 \& }word2)}{\text{prob(word1)}*\text{prob(word2)}}\right)\text{------------------------------------ (1)}$$

Here, prob(word1 & word2) is the probability of word1 and word2 co-occur in the sentence/text document.

Scoring for semantic orientation:

Here, PMI(word1,positive word) and PMI(word1,negative word) calculated, so that we can calculate semantic orientation score.

$$\text{SO Score} = \text{PMI(word1, positive word)} - \text{PMI(word1, negative word)}\text{---------------------- (2)}$$

All phrase value SO Score is calculated and by averaging it we can get average numeric score ,if that is positive then sentence/text document is categorized as positive, if negative then categorized as negative and if value is zero then it should categorized as neutral one[8].

2. Word Based Approach:

Apply stemming procedure on text data for converting derived words to their root by removing end characters and then calculate the polarity of review by comparing the positive and negative word list [9]. If count of positive words is more than negative word in review then review is positive, if less then review is negative otherwise neutral.

3. Hybrid Approach:

We can combine more than one method for calculating sentiment accuracy.PMI based approach in combination with Machine Learning Techniques such as Support Vector Machine, used to calculate accuracy for sentiment polarity for the input review. Similarly, it is possible to combine other approaches.

IV. EXPERIMENTAL ANALYSIS

We have used movie review dataset of 1009 review. In SVM Cross validation, training data and testing data is of 70:30 pattern i.e. 700 review for training data and remaining 409 for testing data. The accuracy value represents the percentage of test texts which were classified correctly by the method. Turney model gives 66% to 68% accuracy value for movie data [8].Here we have taken different combinations of model for analysis.Table-2 shows different model's accuracy. For SO-PMI model, preprocessing is done for removing the unwanted data to minimize the data size. After preprocessing step, selection of phrase value carried out according to pattern discussed in Table-1. Then PMI score is calculated with word such as excellent, good, poor, bad etc. using formula 1 and the SO Score calculated according to formula 2. e.g. $\mathrm{SO\ Score} = \mathrm{PMI}(\mathrm{phrase}, \mathrm{excellent}) - \mathrm{PMI}(\mathrm{phrase}, \mathrm{poor})$ and then assign polarity i.e. positive, negative, neutral to review on the basis of SO score. Using procedure mention in methodology for sentiment analysis in section III assign polarity to review for Word Based Approach, Hybrid(SO-PMI and Word Based Approach), Hybrid SVM(SO-PMI, Word Based Approach and SVM).In Table-2,SO-PMI model has accuracy high at 10 Fold cross validation i.e. 71.72%, Word Based Approach at 5 Fold cross validation i.e. 70.35%, Hybrid(SO-PMI and Word Based Approach) at 3 Fold cross validation i.e.70.91% and Hybrid SVM(SO-PMI, Word Based Approach and SVM) highest accuracy at 10 fold cross validation i.e.71.91%.

Our result is as follows:

Table-2: Analysis using 1009 movie reviews

Model	SVM Cross Validation		
	3 Fold	5 Fold	10 Fold
SO-PMI	69.20	69.32	71.72
Word Based Approach	66.52	70.35	69.52
Hybrid(SO-PMI and Word Based Approach)	70.91	70.12	70.80
Hybrid SVM(SO-PMI, Word Based Approach and SVM)	67.29	70.61	71.91

V. CONCLUSION

As we used the number of positive word and negative word, increased in word list increases the accuracy, so it is very important to choose such a word which will increases the accuracy of the result. The word list completely depends on what type of review is, whether it is movie or any product because many words polarity varies with situation. Turney model gives 66% to 68% accuracy whereas the proposed approach gives 66.52% to 71.91% accuracy. Here accuracy increases due to selection of positive and negative word list for comparing with features as per the review type and another reason is combining the more than one approach i.e. hybrid approach for sentiment analysis.

REFERENCES

1. Umadevi V. ," Sentiment Analysis using Weka" , International Journal of Engineering Trends and Technology , Vol. 18, Issue No. 4, 2014.

2. M. Govindarajan, "Sentiment Analysis of Restaurant Reviews using Hybrid Classification Method", International Journal of Soft Computing and Artificial Intelligence, Vol. 2, Issue No. 1, 2014.

3. M.Sivakumar, C.Karthika, P.Renuga, “A Hybrid Text Classification Approach using KNN And SVM”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Special Issue No. 3, 2014.

4. Rudy Prabowo and Mike Thelwall **,** **“**Sentiment analysis: A Combined Approach**”**, Journal of Informetrics, Vol. 3, Issue No. 2**,** 2009.

5. Noura Farra, Elie Challita, Rawad Abou Assi, Hazem Hajj, “ Sentence-level and Document-level Sentiment Mining for Arabic Texts”, IEEE International Conference on Data Mining workshops, 2010.

6. Gautami Tripathi and Naganna S, “Feature Selection and Classification Approach for Sentiment Analysis”, Machine Learning and Applications: An International Journal , Vol. 2,Issue No. 2, 2015.

7. Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 2002.

8. Mullen, Tony and Nigel Collier, “Sentiment Analysis using Support Vector Machines with Diverse Information Sources”, In Proceedings of EMNLP, 2004.

9. Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", Proceedings of the ACM SIGKDD International Conference on Knowledge, Discovery and Data Mining, 2004.