International Journal of Computer Systems (ISSN: 2394-1065), Volume 02– Issue 04, April, 2015
Available at http://www.ijcsonline.com/

Categorical Difference based Feature Weighting Schemes for Review Analysis

Rashmi Bhutani
Department of Computer Science and Engineering,
Rajasthan Technical University, India

Abstract

Opinion mining is the process to analyse opinions and sentiments to recognize the orientation of people towards different entities. Sentiment analysis attempts to detect the sentiment of a given text whether it is positive or negative. Opinions contain very important information which may be proved to be useful for both customers and organizations. For example, people express pros and cons of different aspects of a product in their opinions, so by analysing them, others can be aware of different aspect of products before buying them. Opinion mining also lets companies improve products, resolve their weaknesses and acquire useful information about their rivals. In this paper, we propose an approach for sentiment analysis. Proposed approach improves the performance of the sentiment analysis by providing improved efficient feature weighting scheme for sentiment analysis.

Keywords: *Sentiment Analysis, Review Analysis, Machine learning Algorithm, Opinion mining.*

I. INTRODUCTION

Online content can be categorized into objective (facts) and opinion (subjective) on the basis of type of data. The main focus of sentiment analysis study is to determine whether a given review document contains positive or negative sentiments (i.e. classify the given review document into negative or positive polarity) [2-6]. Sentiment analysis research can be broadly classified into two methods on the basis of the approach used for classification i.e. Machine learning based approaches and semantic orientation based approaches [10, 12, 14, 15, 17]. Machine learning based methods works in four phases as shown in Figure 1.

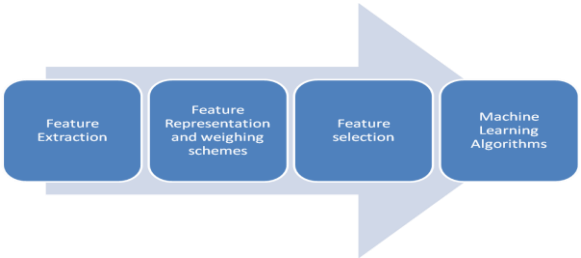

Fig 1: Process for machine learning based sentiment analysis

Whereas semantic orientation based methods basically works in following phases as shown in Figure 2.

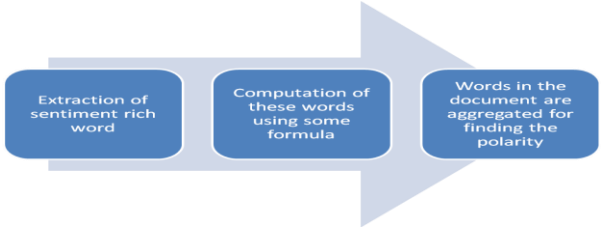

Fig 2: Process for semantic orientation based sentiment analysis

Sentiment analysis can be applied broadly categorised into three levels on the basis of the size of the content. It can be categorised as document level sentiment analysis, sentence level sentiment analysis, and phrase level (aspect level, feature level) sentiment analysis [7, 8, 9].

In this paper, we are emphasizing here on machine learning based approach for document level sentiment analysis. We are investigating the best machine learning algorithm for sentiment analysis and aiming to improve the performance of the sentiment analysis by providing improved efficient feature weighting scheme for sentiment analysis.

Initially, unigrams and bigrams feature vectors are constructed from the unstructured text. Further, various weighting schemes are used to assign the weights to the features which include semantic information and class discriminating ability to determine the importance of the feature for sentiment classification. We use various feature weighting schemes namely Boolean, Term Frequency-Inverse term Frequency (TF-IDF), SentiWordNet (SWN) values, Semantic Orientation (SO) values and Categorical Difference (CD) values. Finally, various machine learning algorithms are used to develop the sentiment analysis model for sentiment classification. We use four machine learning algorithms viz. Support Vector Machine (SVM), Naive Bayes (NB), Adaboost and Bagging algorithms [10-12].

II. PROPOSED APPROACH

In the proposed approach, initially we extract various features i.e. unigrams, bigrams, and then various weighting schemes are used to assign the weights to the extracted features. Finally, we use four classification algorithms to build the machine learning models for sentiment analysis. Here, we present the different weighting schemes used to assign the weights to the features extracted.

2.1Feature Weighting Schemes

Feature weighting schemes are important for assigning the weights to the features. It is important for the classification

results. The most popular feature weighting schemes are described as follows:

2.1.1. Boolean Weighting Scheme

Each document is represented as a vector of words/terms/feature. In binary weighting scheme, feature value is 1 (one) if a term is present otherwise it is 0 (zero).

2.1.2. TF-IDF Weighting scheme

TF-IDF weighting scheme computes the weights as follows.

Weight wij = tfij*idfi,

where tfij is the frequency of term i in document j, and idfi is the inverse document frequency, it measures if a term is common or rare across documents. IDF can be calculated by log (N/F), where N is total number of documents in the corpus, F is number of documents where term i appears. In other words in this method, weights are given to each term according to how rare these terms are in other documents [13-15].

2.1.3. SWN Values

It is a WordNet like lexicon which contain words with three scores as given below i.e.

(i) Positive score

(ii) Negative score, and

(iii) Objective score.

For every word, positive, negative and neutral score are having values between 0.0 to 1.0 and addition of all the score i.e. positive score, negative score and objective score for a word is 1.

2.1.4. Semantic Orientation (SO) Values

Initially, semantic orientation of all the words is calculated using PMI formula. Top n words are selected based on the semantic orientation values of the words. Further, these semantic orientations of words are taken as weights to the feature values.

$$Semanitc\ Orientation(c) = log_2 \frac{P(c, positive)}{P(c, negative)}$$

Here, P(c, positive) is probability of a feature c that it occurs in positive documents

P(c,negative) is the probability that a feature occurs in negative

2.1.5. Computation of Categorical Difference value (CD)

In this method of feature weight scheme, we assign the weights to the features based on their importance depending on their discriminating ability for classification. Initially, the importance of a feature is computed on the basis of the class discriminating ability. Further, these scores are used to assign the weights to the features for the classification. The method to determine the class discriminating ability is as follows.

In the proposed method, a feature is considered as important on the basis of its occurrence in positive or negative labelled document. A feature is important if it is occurring more in only one class and not occurring in other class because if a feature would have this quality that feature would have more discriminating power for identifying a class attribute of a new testing sample. For example, if a word "boring" is occurring only in negative document and not occurring in positive documents then if a new testing document would have a word "boring" that is a strong indication that this new document belongs to negative class.

Algorithm 1: Computation of Categorical Difference value (CD)

Input: Documents (D) with labels (C) negative or positive

Output: All the terms with the CD values, positive and negative polarity with synonym of words and similarity between words

t ← ExtractUniqueTerms(D)

Wp ← DocumentFrequencyInPositiveClass(D,C)

Wn ← DocumentFrequencyInNegativeClass (D,C)

Step-1 Pre-Processing the reviews:

Remove punctuations, stop words:

Remove special symbols

Convert to lower:

Step-2 Get the Feature Vector List:

For w in words:

Replace two or more words

Strip:

If (w in stopwords)

Continue

Else:

Append the file

Return feature vector

Step 3 CD value computation loop

for each t ∈ F

$$CD = \frac{W_p - W_n}{W_p + W_n}$$

end for

Step-4 Training the classifier using the feature vector constructed in step 3

Step 5: Print: sentiment polarity with similar feature words

III. EXPERIMENTS AND RESULTS

A. Dataset Used

We carry out experiments on a large dataset of 50000 movie reviews produced by Maas et al. (2011) [1]. The large dataset IMDB11 contains a training set of 25000 labeled reviews and a test set of 25000 labeled reviews, where training and test sets have 12500 positive reviews and 12500 negative reviews in each. The dataset can be obtained at http://ai.stanford.edu/~amaas/data/sentiment/. In addition, we use product review datasets that consist of amazon product reviews. This benchmark dataset (Blitzer et al. 2007) [16] consists of various domain reviews. We use product reviews of books domains to evaluate the performance of all the proposed methods. This dataset contains 1000 positive and 1000 negative review documents.

B. Evaluation Metrics

F- measure is used for evaluating the performance of the proposed methods. It is based on the precision and recall. Precision for a class C is the fraction of True positives and sum of True Positives (TP) and False Positives (FP). Recall is the fraction of True positives and sum of True Positives and False Negative (FN).

Precision and recall for class Ci can be calculated by:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

Accuracy is commonly used as a measure for categorization techniques.

$$\text{Accuracy}_i = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i}$$

Here,

TPi = Number of documents correctly classified to the class.

FPi = Number of documents incorrectly classified to the class.

FNi = Number of documents not classified to the class.

TNi = Number of documents not classified to the correct class.

$$F_\beta = \frac{(1+\beta^2).\ precision * recall}{(\beta^2 * precision + recall)}$$

Here β is a parameter, which can be used to give the importance to any one precision or recall.

Table 1: Confusion matrix of accuracy evaluation criteria

		Predicted semantic Orientation	
		Positive review	Negative review
Actual sentiment orientation	Positive review	TP	FN
	Negative review	FP	TN

IV. RESULTS AND DISCUSSIONS

Initially, unigrams and bigrams feature vectors are constructed from the unstructured text. Further, various weighting schemes are used which include semantic information namely Boolean, Term Frequency, TF-IDF, SWN values, SO values and CD values. Finally, various machine learning algorithms are used to develop the sentiment analysis model for sentiment classification.

F-measure values of various classifiers with different weighting schemes with unigram features on movie review dataset are given in Table 2. It is clear from the experiments that support vector machines (SVM) classifier performs best among all the machine learning algorithms. Further, it is also clear from the experiments performed that proposed categorical difference (CD) value based feature weighting scheme performs best among all other feature weighting schemes.

Table 2 F-measure (in %) for various classifiers with different weighting schemes with unigram features on Movie review dataset

S. No.	**Feature Weighting Scheme**	**SVM**	**NB**	**Adaboost**	**Bagging**
1	Boolean	83.6	81.2	66.8	78.1
2	TF-IDF	81.6	80.2	65.4	75.4
3	SWN values	76.2	74.2	63.5	71.2
4	SO values	84.8	82.8	68.9	80.1
5	CD values	85.7	84.3	70.2	81.4

Similarly, the F-measure values are reported for various classifiers with different weighting schemes with bigram features on Movie review dataset in Table 3. It is also clear from these experiments that support vector machines (SVM) classifier performs best among all the machine learning algorithms for bigram features and proposed categorical difference (CD) value based feature weighting scheme performs best among all other feature weighting schemes.

Table 3 F-measure (in %) for various classifiers with different weighting schemes with bigram features on Movie review dataset

S. No.	Feature Weighting Scheme	SVM	NB	Adaboost	Bagging
1	Boolean	77.2	74.2	63.8	75.2
2	Term frequency	75.4	73.3	62.4	74.3
3	SO values	78.7	76.7	64.9	76.4
4	CD values	79.8	78.3	67.3	77.5

Further, support vector machine (SVM) classifier performs best for bigram features among all the machine learning algorithms as shown in Table 3. For example, support vector machine (SVM) classifier gives 77.2 % with Boolean feature weighting scheme which is better than other classifiers viz. Naive bayes, Adaboost, and boosting algorithms.

To make the experiments more stable, we experiment with various proposed methods with Book review dataset. Table 4 presents F-measure (in %) values for various classifiers with different weighting schemes with unigram features on Book review dataset. For book review datasets, it is clear from the Table 4 that categorical difference based feature weighting scheme performs best among other feature weighting schemes

Table 4 F-measure (in %) for various classifiers with different weighting schemes with unigram features on Book review dataset

S. No.	Feature Weighting Scheme	SVM	NB	Adaboost	Bagging
1	Boolean	76.2	75.8	70.1	72.8
2	TF-IDF	75.9	74.5	68.7	70.7
3	SWN values	72.1	71.9	67.1	68.9
4	SO values	78.9	77.8	72.9	74.7
5	CD values	80.2	79.2	75.2	76.6

Similarly, Table 5 presents F-measure (in %) for various classifiers with different weighting schemes with bigram features on Book review dataset. It is shown with the experimental results that support vector machine (SVM) classifier gives 72.1 % with Boolean feature weighting scheme with bigram features which is better than other classifiers Further, proposed semantic orientation based feature weighting scheme produced better results as compared to Boolean, TF-IDF and SWN values based feature weighting methods.

Table 5 F-measure (in %) for various classifiers with different weighting schemes with bigram features on Book review dataset

S. No.	Feature Weighting Scheme	SVM	NB	Adaboost	Bagging
1	Boolean	72.1	71.9	67.2	70.3
2	TF-IDF	70.3	69.4	65.6	69.1
3	SO values	74.5	72.4	68.5	72.5
4	CD values	75.3	74.3	70.3	73.8

Unigrams and bigrams feature vectors are constructed from the unstructured text. Further, various weighting schemes are used which include semantic information namely Boolean, Term Frequency, TF-IDF, SWN values, SO values and CD values. Finally, various machine learning algorithms are used to develop the sentiment analysis model for sentiment classification. Experimental results show that proposed categorical difference (CD) value based feature weighting schemes performed best among other feature weighting schemes. Further, among various machine learning algorithms, SVM classifier performs better than other algorithms.

V. CONCLUSION AND FUTURE WORK

Opinions contain worthy information that are useful for both customers and organizations. In this paper, we propose a categorical difference (CD) value based feature weighting scheme which performed better than other feature weighting schemes. Further, support vector machine (SVM) classifier outperformed other machine learning algorithms. In future, we would try to explore more effective feature weighting scheme which include more information in assigning the weights to the features. In addition, we would like to do the analysis more fine grained by looking at the each aspect of the review. For example, a camera may be good for a set of people as they like „size" and „battery" of the camera, simultaneously the same camera may not be liked by other set of people due to expensive feature of the camera

VI. REFERENCES

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, "Learning word vectors for sentiment analysis", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol: 1, No:1, pages 142–150, 2011.

[2] B.Agarwal, N. Mittal, "Text Classification using Machine Learning Methods- A Survey", In 2nd International Conference on Soft Computing for Problem Solving (SocPros-2012), Vol:236, pp: 701-710, 2012.

[3] B.Agarwal, N. Mittal, "Sentiment Classification using Rough Set based Hybrid Feature Selection", InProceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA"13), pages 115–119 , NAACL-HLT, Atlanta, USA, 2013.

[4] Justin Martineau and Tim Finin, "Delta tfidf: an improved feature space for sentiment analysis", In Proceedings of the Third Annual Conference on Weblogs and Social Media, pages 258–261, 2009.

[5] Agarwal B., Sharma V. K., Mittal N. "Sentiment Classification of Review Documents using Phrase Patterns", International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages: 1577-1580, 2013.

[6] N. Mittal, B. Agarwal, S.Laddha, M. Sharma, "Aspect Based Analysis for Rating Prediction of the Restaurant Reviews", In International Journal of Computer Systems, pages: 59-62, Vol. 02, Issue: 03, March, 2015.

[7] Dai L., Chen H., Li X., "Improving Sentiment Classification using Feature Highlighting and Feature Bagging". In Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW), pages: 61-66, 2011.

[8] Pang B., Lee L., Vaithyanathan S., "Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language.

[9] Basant Agarwal, Namita Mittal, "Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification", In Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), COLING 2012, pages 17–26, 2012.

[10] Liu B., "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012.

[11] Dang Y., Zhang Y., Chen H., "A lexicon Enhanced method for sentiment classification: An experiment on Online Product Reviews", In IEEE Intelligent System, Vol: 25, No: 4, pages: 46-53, 2010.

[12] Deng Z.H., Luo K.H., Yu H.L., "A study of supervised term weighting scheme for sentiment analysis", In Expert Systems with Applications, Vol: 41, No: 7, pages: 3506-3513, 2014.

[13] B. Agarwal, N. Mittal, "Prominent Feature Extraction for Review Analysis: An Empirical Study", In Journal of Experimental and theoretical Artificial Intelligence, Taylor Francis, 2014, DOI:10.1080/0952813X.2014.977830.

[14] B. Agarwal, N. Mittal, "Semantic Feature Clustering for Sentiment Analysis of English Reviews", In IETE Journal of Research, Taylor Francis, Vol: 60 (6), pages 414-422, 2014.

[15] B. Agarwal, N. Mittal, P. Bansal, S. Garg, "Sentiment Analysis Using Common-Sense and Context Information", In Computational Intelligence and Neuroscience, 9 pages, DOI: http://dx.doi.org/10.1155/2015/715730, Article ID: 715730, 2015.

[16] Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 440–447 (2007)

[17] B. Agarwal, N. Mittal, "Optimal Feature Selection for Sentiment Analysis", In 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013),Vol-7817, pages-13-24, Greece, Samos. 2013.