Subjectivity detection and Semantic orientation based Methods for Sentiment Analysis

Sandhya Khanna, Savita Shiwani

Abstract— Sentiment Analysis research deals with the extraction of the opinion expressed by people about specific topic from the text review documents. Sentiment analysis research has increased tremendously in recent time. Main challenge with movie review analysis is that real facts and discussion of plot are mixed in the sentiment expressing sentences in the review. In this paper, two methods are investigated to eliminate the objective sentences for better classification results of movie review analysis. In addition, performance of unigrams and rule based phrases are investigated for sentiment analysis. All the experiments are performed on the benchmark dataset i.e. Cornell's movie review dataset. Experimental results show that Subjectivity detection i.e. elimination of objective sentences improves the performance of sentiment analysis and phrases performs better than unigrams for sentiment analysis.

Index Terms—Sentiment analysis, Subjectivity detection, semantic orientation, Naive bayes, Polarity detection, Phrase extraction,Sentiwordnet.

1. INTRODUCTION

Sentiment Analysis research deals with the extraction of the opinion expressed by people about specific topic from the text review documents [2]. Sentiment analysis research has increased tremendously in recent time. It is very important for users as well as e-commerce companies as users may use the online reviews for their purchasing decisions and companies can use these reviews for improving their products. For example, whenever user wants to purchase mobile phone, he refers the online reviews related to that mobile or reviews related to other companies mobiles phones and compares the phones on the basis of experiences of different users written in the reviews, however, he may online refer some of the reviews. Therefore it is required to have a system which can analysis thousands of reviews and can provide an aggregate opinion of users about any product like mobiles, automobile etc. In addition, sentiment analysis can be important in political or health fields as political parties may be interested in the opinion of common people about their policies or about their party, and in medical field it may be important to know that which doctor is specialized or which medicine has been proved successful on the basis of reviews written by various patients.

Main challenge with the movie review analysis is that real facts are mixed with other objective sentences like discussions about the plot in the movie, discussion about the characteristics of the actors and actress etc. with the opinion sentences. These objective sentences are not important for identifying the opinion expressed in the review. There are mainly two types of approaches for sentiment analysis i.e. Machine learning algorithms based approaches [1], [7], [9] and semantic orientation based approaches [3]. This paper focus on the semantic orientation based approaches for sentiment analysis. Semantic orientation based approaches mainly works as follows. Initially, sentiment-rich features are extracted from the review document, and then semantic orientation of these features are computed using some formula and finally overall semantic orientation or polarity is determined by aggregating the semantic orientations of all the words in the document.

Objective of this paper is two-fold. First is to explore the best features that conveys better sentiments and second is to investigate a method which can improve the performance of the sentiment classification by eliminating the objective sentences. Therefore, initially experiments are performed to explore which feature are best i.e. unigrams or phrases. And, in further experiments two subjectivity detection methods are employed to know which method is best. Experimental results show that phrases are better in capturing the sentiment from the documents it is due to the fact that phrases can incorporate the contextual information unlike unigrams. Next, by adding subjectivity detection into simple semantic orientation based methods for sentiment analysis increases the performance. Further, SentiWordNet based method performs better than naïve bayes based subjectivity detection method with adjectives, adverbs, nouns and verbs.

The organization of this paper is as follows. Section 2 describes the related work. Proposed methodology is described in Section 3. Experimental results and discussion are presented in Section 4, and finally Section 5 concludes the paper with discussion of possible future work.

2. RELATED WORK

Semantic orientation based approaches has been explored for Sentiment analysis in the literature. In [3], authors propose an unsupervised method for identifying the polarity of a movie review document. Initially, they extracted two-word phrases using fixed POS based patterns, then semantic orientation of those phrases are

computed using Point-wise Mutual Information (PMI) method. Finally, overall polarity of the document is recognized by aggregating the semantic orientation of all the phrases.In [12], authors proposed the method for automatically identification of POS based pattern for extraction of polar phrases. They applied various feature selection methods namely Information Gain (IG), Chi-Squares (CHI), and Document Frequency (DF) for identification of important phrase patterns.Further, they applied PMI method for calculation of semantic orientation of phrases. In [8], authors constructed some phrase patterns with adjectives, adverbs, prepositions, conjunctions, noun, and verbs. Further, semantic orientation of these phrases is calculated using unsupervised method. In [11], authors use post-processing approach for phrase extraction for Japanese documents; they first extract n-gram feature and counting their frequency and remove n-grams of lower frequency. Further select rule based features from n-grams. Then recalculate the frequency of features.

Authors in [5], applied min-cut method for subjective sentence extraction, and further classification algorithms like Naive Bayes (NB), Support Vector Machine (SVM) are applied for polarity classification.In [10],, authors used WordNet based method for the effective incorporation of linguistic information for subjective sentence extraction, further they also usedSupport Vector Machine classifier for polarity classification.Authors in [4],extracted polar sentences from the Japanese HTML documents using language structural clues. Next, phrases are extracted from polar sentences using dependency parser. Further semantic score of each polar phrase is computed using Chi-square and PMI method.

3. PROPOSED APPROACH

Work flow of the proposed approach is presented in Figure 1. Initially dataset are pre-processed by Part-of-Speech (POS) tagging, Further, negation handling is applied. Then subjective sentences are selected using two methods based of SentiWordNet and Naïve Bayes algorithm i.e. discarding objective sentences from the documents. Further, one-word and two-word features are extracted i.e. POS based unigrams and POS based fixed patterns.

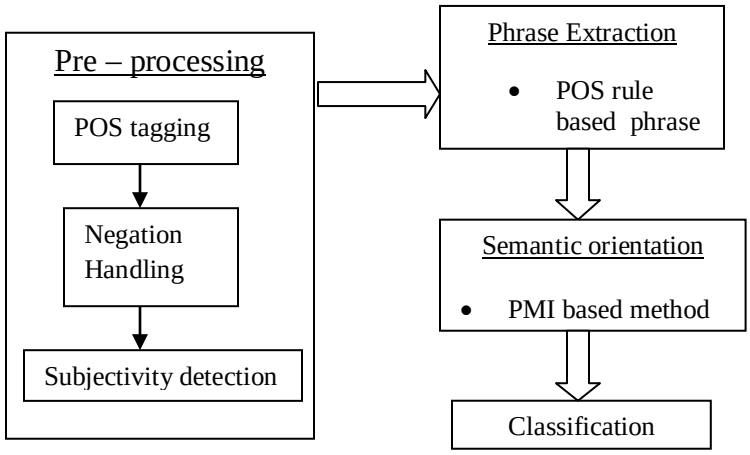

Figure 1: Proposed Approach

After extraction of sentiment-rich phrases, semantic orientations of all these phrases are computed using Point-wise Mutual Information (PMI) method. Unigrams are extracted using POS i.e. adjectives, adverbs, verbs etc. and semantic orientations of unigrams arealso computed using PMI method and with the help of SentiWordNet (SWN). Finally, overall semantic orientation of the document is determined by aggregating the semantic orientation of all the phrases and unigrams in the document.

3.1 SUBJECTIVITY DETECTION:

Subjective sentences are extracted from documents because mostly movie review documents contain two types of sentences, one which talks about the actors or plot in the movie and other which express the sentiment about the movie. Sentiment analysis and opinion mining task is interested in the "about the movie" part of review, and the sentences which express "about the movie" part are called subjective sentence. Generally, people express "about the movie" part by strong adjectives. For example: "there is a great deal of corny dialogue and preposterous moments" contain adjectives great, corny and preposterous. In this paper, two methods are explored for determining if a given sentence is a subjective sentence or objective sentence i.e. Naïve Bayes based subjectivity detector and SentiWordNet.

3.1.1 NAÏVE BAYES CLASSIFIER BASED SUBJECTIVITY DETECTOR

To determine if a given sentence is subjective or objective, a learning model is developed based on Naive Bayes classifier on the subjective dataset. Subjective dataset contains 5000 subjective and 5000 Objective sentences(http://www.cs.cornell.edu/people/pabo/movie-review-data/). This dataset is used for subjectivity classification. Naïve bayes classifier initially computes the probability that a given instance belongs to which class, and then it labels the instance whose probability is highest. Similarly, in this method, probability that a give sentence belongs to subjective class or objective class is computed and further sentence is labeled according to which probability (probability that given sentence is subjective or objective) is high. Process to determine the probability that a given sentence belongs to subjective or objective class is determined in two phases, in the first phase a lexicon is build which has subjective and objective score of words computed using subjectivity dataset. And in the second phase, for every sentence of testing movie review, subjective and objective score is retrieved from the lexicon and further average subjective and objective score of all the words in the sentence are computed, finally based on these average score, it is determined that a given sentence is subjective or objective. Detailed description of this process is described in following subsections.

1. Initially, adjectives/adverbs are extracted from subjective dataset using POS tagger and frequency of adjective/adverb in subjective and objective sentences i.e. f(w,sub) and f(w,obj) are computed respectively. Further, probability that a given sentence belongs to subjective sentences p(w,sub) is computed as given in Equation (1) and similarly probability that a given sentence belongs to objective sentences p(w,obj) is computed as given in Equation (2), and built a lexicon which contain adjectives/adverb and their subjective and objective probability.

$$p(w,sub) = \frac{f(w,sub)}{f(w,sub)+f(w,obj)} \quad (1)$$

$$p(w,obj) = \frac{f(w,obj)}{f(w,sub)+f(w,obj)}$$

2. For all the sentences in the testing review documents POS tagging is performed. Then adjectives/adverbs are extracted from a sentence of a document. Next, subjective and objective score are retrieved for all these words from lexicon build in first phase. Finally, average subjective and objective scores of all adjective/adverb of a sentence are computed using Equation (3), (4).

$$Sub(s) = \frac{\sum_i p(w_i,sub)}{n}$$

$$Obj(s) = \frac{\sum_i p(w_i,obj)}{n}$$

For each sentence subjective Sub(s) and objective Obj(s) scores are computed, and classify the sentence into objective by two ways.

(1) If Sub(s)>Obj(s) then sentence is considered as subjective else sentence is objective and discarded the objective sentence.

(2) Sort the sentences by their average subjective score and select top 80% or 85% of sentences.

3.1.2 SENTIWORDNET METHOD

SentiWordNet based method works retrieves the polarity scores for each word from the standard polarity lexicon. Subjective and objective scores of each sentence are computed using Equation (5) and (6) with the help of SentiWordNet [6]. Various methods are explored to compute the subjective and objective score of a sentence to investigate the performance of only adjectives, adjectives with adverbs and adjectives with adverbs, noun, and verbs. Since, adjectives are intuitively the most important in expressing sentiments; therefore, the more weights are given to the adjectives as compared to other part-of-speech words. For example, "this is a very nice movie". In this example, "nice" is the adjective which is the most important word in this sentence that is conveying information that a given sentence is expressing some sentiment and is likely to be a subjective sentence.

If a word is an adjective then subjective and objective scores are computed using Equation (5) and (6).

$$sub(w_i) = \alpha * (|pos(w_i)| + |neg(w_i)|)$$

$$obj(w_i) = \alpha * \left(1 - sub(w_i)\right) \quad (2)$$

And if a word is an adverb, verb, noun then subjective and objective scores arecomputed using Equations (7) and (8).

$$sub(w_i) = \beta * (|pos(w_i)| + |neg(w_i)|) \quad (7)$$

$$obj(w_i) = \beta * (1 - sub(w_i)) \quad (8)$$

Here pos(wi) is a positive score, and neg(wi) is a negative score of a given wordretrieved from SentiWordNet. α and β are constants where α > β, in our experiment we take α = 2 and β = 0.5 . If a sentence contains n words then subjective and objective scoresof the sentence are computed using Equations (9) and (10).

$$sub_{score} = \frac{\sum_{i=(1,n)} sub(w_i)}{n} \quad (9)$$

$$obj_{score} = \frac{\sum_{i=(1,n)} obj(w_i)}{n} \quad (10)$$

After determining the average subjective and objective scores for the sentences, two cases are taken for further processing.

(1) If subscore>objscore then sentence is considered subjective else considered as objective and then objective sentence are discarded.

(2) Sort the sentences by their average subjective score and select top 80% or 85% of sentences.

3.2 Feature Extraction

In the proposed approach, two types of features are extracted i.e. one-word features and two-word features called unigrams and phrases respectively.

3.2.1. Unigrams

In unigram as features adjectives, adverbs, nouns and verbs part-of-speeches are considered as mostly these part-of-speech words convey the sentiments.

3.2.1 Phrases

Phrases are very essential for extraction of contextual information which is very important for sentiment analysis. Phrases are capable of capturing contextual information like “not good”, “unpredictable story”, “amazing movie” etc. POS pattern based phrases are extracted that follow the fixed pattern as given in Table 1. POS based rules are able to extract sentiment-rich phrases which incorporates contextual information from the text.

Table 1.Rules for extracting phrases

S.no.	First Word	Second Word
1	Adjective	Noun
2	Adverb	Adjective
3	Adjective	Adjective
4	Noun	Adjective
5	Adverb	Verb

3.3. Semantic Orientation

After extraction of features, semantic orientation of each feature is determined using Point-wise Mutual Information (PMI) method. It is generally used to calculate the strength of association between a phrase and positive or negative sentences. Semantic orientation of a word or phrase c can be defined [4] as Equation 1.

$$SO(c) = log_2 \frac{P(c,pos)}{P(c,neg)} \quad \dots (1)$$

Here, P(c,pos) is probability of a phrase or word that this feature occurred in how many number of positive documents divided by total number of positive documents. Similarly, P(c,neg)is the probability that a phrase or word occurs in how many of negative document divided by total number of negative documents.

3.4 Semantic Orientation Aggregation

After determination of semantic orientation of each feature using Point-wise Mutual Information (PMI) in the document, overall positive or negative semantic orientation of the document is determined by summing up the semantic orientation of all the words in the document. Finally, overall semantic orientation of the document is determined as positive if the aggregated semantic orientation value of the document is greater than zero else it is determined as negative.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate the proposed approach, publically available dataset is used i.e. Movie review dataset provided byCornell university [5]. This dataset contains 2000 movie reviews containing 1000 positive and 1000 negative reviews (http://www.cs.cornell.edu/people/pabo/movie-review-data/). For all the experiments, 700 review documents from each class are randomly selected for training and remaining 300 review documents from each class are used for testing the propose approach for sentiment analysis. To evaluate the performance of the proposed method accuracy is computed by Equation 2.

$$Accuracy = \frac{total\ number\ of\ correctly\ classi}{Total\ number\ of\ testing\ s} \quad ..(2)$$

All the experiments performed in the proposed approach for sentiment classification can be categorised into two experiments, objective of the first experiment is to investigate the best feature i.e. unigrams or phrases for sentiment classification. Therefore, subjectivity detection is not applied and whole document is considered for classification of the given review document into positive or negative polar document. The main objective of the second experiment is to investigate the impact of elimination of objective sentences from the review document. Two methods are applied for identification of subjective sentences as discussed in previous chapters.

EXPERIMENT 1: Initially, Stanford POS tagger is applied on document corpus; further negation handling is applied on corpus. Further, without applying any subjectivity detection two types of features are extracted i.e. unigrams and POS pattern based phrases. In case of unigrams, only those unigrams are extracted which are POS tagged by adjective (JJ), adverb (RB), verb (VB) and noun (NN).And, phrases are extracted which conform to the predefined patterns as given in Table 1. Further, semantic scoresof these POS based unigrams and phrases are computed by using Equation(1). Further, lexicons of unigrams words and phrases with their semantic orientations are built. Now, at the time of testing a review document, semantic orientations of all the POS tagged unigrams are retrieved from this lexicon. Finally, the semantic orientations of all the words of the testing documents are aggregated for determining the polarity of the overall document as positive or negative. Similarly, overall semantic orientation of the document is determined with the POS pattern based phrases. Experimental results for the effectiveness of the unigrams and phrases for the sentiment analysis are shown in Table 2.

Table 2: Accuracy for unigrams and phrases for movie review dataset

	Positively correctly classified	Negatively correctly classified	Total correctly classified	Accuracy In (%)
Unigrams	277	163	440	73.33
Phrases	248	209	457	76.17

Experimental results for unigrams and phrases for sentiment analysis are shown in Table 2. Experiments performed only on unigrams which are POS tagged by JJ/RB/VB/NN and it correctly classify 277 (300) positive and 163 (300) negative documents and accuracy computed based on only unigram is 73.33 %. With phrases are features accuracy is increased from 73.33% to 76.17%. From the experimental results, it is clear that phrases are better than unigrams for sentiment classification.

EXPERIMENT 2: Subjectivity detection is performed to evaluate impact of subjectivity detection on sentiment classification. First of all objective sentence are discarded using both basic subjectivity detector and SentiWordNet method. Further, phrases are extracted using POS based rules as given in Table 1. Next, semantic orientations of all these phrases are computed using Equation (1) and built the phrase lexicon. Further, at the time of classifying a new test document all the phrases are extracted with POS based rules, and then overall polarity is determined by averaging semantic scores of all the phrases in the document. To investigate the effect of elimination of objective sentences, different numbers of sentences are eliminated by considering 80%, 85% and 90% of sentences.

Table 3: Accuracy of phrase with subjectivity detection for movie review dataset

	Sub(s) >Obj(s)	Top 80%	Top 85%	Top 90%
Naïve bayes based Subjectivity detector	77.33%	73.00%	75.17%	75.83%
SentiWordNet based on Adj.	76.00%	74.33%	75.33%	77.13%
SentiWordNet based on Adj./Adv.	76.00%	77.00%	77.83%	77.83%
SentiWordNet based on Adj./Adv./Verb/ Noun	76.33%	78.00%	77.17%	77.17%

Accuracies for both types of subjectivity detection methods with various setting of experiments are reported in Table 3 and also graphically shown in Figure 1.

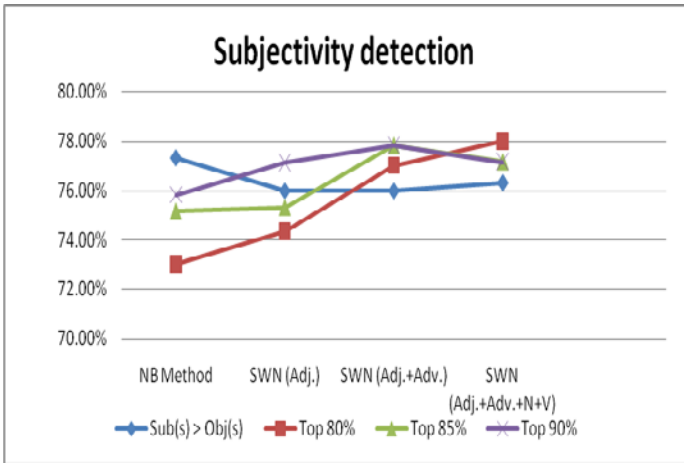

Figure 1 Accuracies for subjectivity detection

Experimental results show that phrase without subjectivity detection shows an accuracy of 76.17%, and accuracy after subjectivity detection by different method reported in Table 3, basic subjectivity detector give an accuracy improvement of 77.33% (+1.52%) over 76.17% and by SentiWordNet method accuracy improves up to 78% (+2.40%) over 76.17%. Experimental results show that SentiWordNet method of subjectivity detection performs better than Naïve bayes based method for sentiment classification.

5. Conclusion and Future Work

Sentiment Analysis research deals with the extraction of the opinion expressed by people about specific topic from the text review documents. Objective of this thesis is two-fold. First is to explore the best features that conveys better sentiments and second is to investigate a method which can improve the performance of the sentiment classification by eliminating the objective sentences. Therefore, initially experiments are performed to explore which feature are best i.e. unigrams or phrases. And, in further experiments two subjectivity detection methods are employed to know which method is best. Experimental results show that phrases are better in capturing the sentiment from the documents it is due to the fact that phrases can incorporate the contextual information unlike unigrams. Next, by adding subjectivity detection into simple semantic orientation based methods for sentiment analysis increases the performance. Further, SentiWordNet based method performs better than naïve bayes based subjectivity detection method with adjectives, adverbs, nouns and verbs.

In future, more various sophisticated features can be explored that can incorporate more syntactic and long distance relation among words in the sentences, because these type of feature can be useful for sentiment analysis. Further, the proposed methods may be explored on various other datasets with different domains. In addition, the proposed methods may be tested on reviews written in non-English language.

References

1. [1] B. Pang, L. Lee, "thumbs up? Sentiment classification using machine learning technique", EMNLP 2002, pages. 79-86.
2. [2] B. Liu, "Sentiment Analysis and Opinion Mining". Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers 2012.
3. [3] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" ,ACL 2002, pages.417-424.
4. [4] N. Kaji, M. Kitsuregawa, "Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents", ACL 2007, pages. 1075-1083.
5. [5] B. Pang, L. Lee ,"Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts",2005.
6. [6] A. Das, S. Bandyopadhyay, "Subjectivity detection in English and Bengali : A CRF-based Approach". In 7th International conference on NLP (ICON). 2009.
7. [7] B. Agarwal, N. Mittal, "Optimal Feature Selection Methods for Sentiment Analysis", In 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013),Vol-7817, pages-13-24, 2013.
8. [8] Z. Fei, J. Liu, G. Wu, "Sentiment classification using phrase pattern", IEEE,2004.
9. [9] B. Agarwal, N. 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. [10] P. Bhattacharyya, S. Verma, "Incorporating semantic knowledge for sentiment analysis", ICON 2008.
11. [11] Y. Okuno, "Phrase extraction for Japanese predictive input method as post-processing", In WTIM, 2011, pages. 48-52.
12. [12] R.Mukras, N. Wiratunga, R. Lothian. "Selecting Bi-Tags for Sentiment Analysis of Text", In the Proceedings of AI-2007, 27th SGAI IntConf on innovative techniques and applications of AI, pages 181–194, 2007.