Feature based Star Rating of Reviews: A Knowledge-Based Approach for Document Sentiment Classification

Shaishav Agrawal[1] and Tanveer J. Siddiqui[2]

[1]Indian Institute of Information Technology Allahabad, India
[2]Department of Electronics and Communication and Computer Science, University of Allahabad, India

[1]shaishav.engr@gmail.com, [2]jktanveer@yahoo.com

Abstract

This paper presents a novel knowledge-based approach for star rating of reviews. It uses SentiWordNet and linguistic heuristics to determine sentiment orientation of sentences, which is used to assign a positive, negative and objective score to document to achieve 5-star rating of movie reviews. A method for generating ratings based on individual features is also presented. The experimental results on sentiment scale dataset demonstrate the effectiveness of our approach.

Keywords: *Opinion Mining, Sentiment Polarity, Sentiment Classification, Star Rating, SentiWordNet.*

1. Introduction

Quite often people base their decision on "what other people think". Earlier people used to seek opinions of others on "which refrigerator is good?", "Whom to vote in elections?" etc. Now, World Wide Web provides an alternative and rich source of such information. According to a survey conducted in USA, 81% of Internet users do online research before purchasing many products [12]. Blogs, customer review sites, forums, provide a platform where people communicate their experiences, views and emotions on all kinds of topics. The sentiments expressed in these documents can be classified as positive and negative, or into an n-point scale, e.g., very good, good, satisfactory, bad, very bad. Websites such as *epinion.com, carwale.com, ebay.com, imdb.com* collect reviews from users on consumer products, automobiles, computer accessories, movies, etc; and make use of human experts to rate them according to pros & cons, good & bad, etc. The ever increasing size of opinionated content has led to the development of automatic sentiment classification methods.

A large body of research on sentiment classification attempt to classify word, sentence or document as positive or negative on the basis of sentiment. Most of these works make use of polar terms such as "good", "bad", "fantastic", "worst", etc. to determine polarity of sentence or document. We argue that the context of sentence plays an important role in deciding its semantic orientation than individual words. Consider the following sentence:

The movie is not good.

Although the word "good" gives a positive expression, the negation word "not" transforms it into negative one. Kennedy and Inkpen[14] used negation rules to carry out negative transformation. However, presence of negation word in context is not single deciding factor for such transformation. Words like long, short, rough, shiny, etc. can express both positive and negative opinions depending on the context. Consider following sentences:

The movie was very long. I felt bored.

I am using this shampoo from last six months. This keeps my hair long, shiny and strong.

The same word “long” expresses negative sentiment in the first example while positive in the second. The correct sentiment of “long” in these sentences can be judged only by considering the context in which it occurs. We propose the use of linguistic heuristics to consider the context in the classification process. In this paper we propose a new knowledge-based approach for classifying reviews. The proposed approach identifies correct sentiments with the help of linguistic heuristics. Unlike existing unsupervised and knowledge-based approaches for document sentiment classification which focus on classifying documents into positive and negative, we determine the strength of opinions on a five point scale (star rating). The proposed method uses SentiWordNet 1.0[1]; a lexical resource designed by Esuli and Sebastiani[8] which contains normalized positive and negative scores of synsets taken from WordNet 2.0[2] for different part of speech and senses. We have also proposed a method to classify the sentiments and generate rating of the document based on individual features. The motivation behind this work is that most of the people want to know the rating of any product, movie, or service according to particular features. This type of rating may be totally different from the overall rating and can change the mind of any consumer to choose a product or service. Suppose a person wants to buy a laptop of a company offering best after sale service and can compromise with other features. Then the rating based on service feature will be more useful to him/her than overall rating. The present work extends the work reported in Agrawal and Siddiqui[1]. The work in Agrawal and Siddiqui[1] focuses on identifying sentiment polarity (positive or negative) whereas the present work focuses on star rating problem. We observed a maximum accuracy of 84.6% on the dataset used by Pang and Lee[18] which is better than earlier reported results on the same dataset.

The rest of the paper is organized as follows. Section 2 describes related work on opinion finding and ranking. The proposed methods including feature selection, scoring methods, repair heuristics and ranking strategies are discussed in Section 3. Section 4 regards our experiments. It describes the dataset, the evaluation measures and the results. Finally, Section 5 concludes the paper and gives ideas for future research direction.

2. Related Work

Opinion mining is relatively a new area of research in the field of natural language processing. A number of approaches exist in literature that attempt to classify word, sentence or document as positive or negative on the basis of sentiments being expressed. Most of these works use word as the basic sentiment bearing unit. There are two main tasks at word level sentiment analysis namely, subjectivity analysis and orientation detection. Subjectivity analysis is concerned with identifying whether a word is subjective or objective, whereas orientation detection is concerned with

[1] Available at, http://sentiwordnet.isti.cnr.it/

[2] WordNet 2.0, http://wordnetcode.princeton.edu/2.0/WordNet-2.0.tar.gz.

identification of semantic orientation of a subjective term, i.e. whether it expresses a positive opinion or negative opinion. A large body of research work focuses on identifying semantic orientation of individual words or phrases. Sentence or document level sentiment classification is achieved by combining word or phrase level sentiment information. Existing approaches to sentiment analysis for text can be broadly categorized into symbolic and machine learning approaches. The symbolic approach [11, 23, 15, 13] uses manually crafted rules and lexicons, whereas the machine learning approach [22, 16, 7, 9, 6, 8, 3, 21] uses supervised or weakly supervised learning to construct a model from a large training corpus.

Symbolic approaches consider document as a bag of words and attempt to identify sentiment by aggregating sentiment of words appearing in it. The work in Hatzivassiloglou and McKeown [11] and Kamps, et. al., [13] focuses on adjectives only. Hatzivassiloglou and McKeown[11] argued that adjectives are strong predictors of sentiments. The work in Kamps, et. al., [13] involves construction of a graph between two seed terms “good” and “bad” and target adjectives using WordNet synonymy relation. However, other parts of speech may also convey sentiments. Published literature focusing on nouns and verbs include Riloff, et. al., [20] and Kim and Hovy [15]. The work in Riloff, et. al., [20] focuses on extracting subjective nouns. Kim and Hovy [15] used WordNet semantic relations for word level sentiment classification. They assembled a small amount of positive and negative seed terms and expanded it using synonym and antonym relation from WordNet. Seed words include adjectives and verbs. For verbs only synonym relation was used. The underlying assumption was that synonyms of positive words are positive and antonyms mostly negative, and vice versa. The polarity of sentiment bearing words was combined to produce sentiment of the whole sentence. Turney and Littman [23] determine semantic orientation of phrase by calculating Point wise Mutual Information (PMI) value with pre-defined positive and negative seed sets. They used average of sentiment information of phrases to predict sentiment orientation of a review. Baroni and Vegnaduzzo [2] used the PMI Method to determine term subjectivity. Instead of using word frequency like Turney and Littman[23] they used probabilities.

Machine learning approaches have also been proposed for sentiment classification task. Pang and Lee [16] defined a novel machine learning approach for the sentiment polarity analysis of sentences by applying text categorization techniques to find minimum cuts in the graph. Before this Pang, et. al., [19] used Support vector Machines (SVMs) as a default polarity classifier for polarity analysis while Turney [22] presented a simple unsupervised learning algorithm for classifying the reviews as Positive or Negative in sentiments. Gamon, et. al., [9] used the combination of clustering algorithm and a machine learned sentiment classifier to determine the opinion from the customer reviews. Esuli and Sebastiani [7] introduced a semi-supervised learning method for determining term orientation. They used glosses as the textual definitions of terms in WordNet and later proposed a variant of it [6]. They took the idea further to design a lexical resource called SentiWordNet 1.0[3] [8] in which each WordNet synset is associated to three numerical values in the range 0.0 to 1.0, describing the strength of objectivity, positivity, and negativity of the terms. SentiWordNet has been used for determining sentiment orientation of documents [5, 4, 1]. A comprehensive review of various techniques for sentiment analysis can be found in Pang and Lee [17].

[3] Available at, http://sentiwordnet.isti.cnr.it/.

Instead of classifying sentences and documents as positive and negative some work attempts to identify subjectivity degree and polarity degree, also known as star rating of reviews [18]. Polarity degree can be classified as weakly positive or negative, mildly positive or negative, and strongly positive or negative or can be in the form of some normalized score or in some rating system (4 Star, 5 Star etc). Wilson, et. al., [24] first presented experimental results to classify the strengths of opinion. They measured subjectivity degree in terms of four categories: neutral, low, medium, and high. Pang and Lee [18] determined the strength of opinion on a three points and four points scale i.e. one star to four star rating. They first evaluated human performance, and then applied a Meta algorithm which labeled the output of n-ary classifier in such a way that items in one category received similar labels by using 'metric label formulation'. Ghose, et. al., [10] have used an entirely different rating strategy; they have used dollar value (financial cost) of features to determine the positivity and negativity of expressions. The work presented in this paper focuses on star rating problem. We propose a novel knowledge-based approach to the star rating problem and evaluate it on the dataset used in Pang and Lee [18].

3. Methodology

We propose and evaluate three different approaches for rating generation of reviews. These methods can be used for rating generation based on individual features as discussed in Section 3.4.

3.1 SentiWordNet Average Scoring Approach

In this approach the score of each term is calculated as the average score of all synsets[4] of that term. The steps in this approach are discussed below:

Step 1: Preprocessing

In the preprocessing step, first the sentence boundary is identified and then the text is tokenized. Extra white spaces, html tags, new lines and unrelated extra characters and special symbols are removed. Stop words are also removed as they do not belong to any of the four parts of speech (Noun, Adjective, Verb, and Adverb) present in the SentiWordNet and they do not affect the opinion expressed in the document. The list of stop words used in this work excludes adverbs like very, more etc. and conjunctions such as and, but, etc. which can affect the subjective information of text. SentiWordNet is also preprocessed to remove the POS tag and sense number from synset terms and to stem verb synsets.

Step 2: Word Scoring

Each word in the document that appears in the SentiWordNet is assigned a positive, negative and objective score. The positive score is calculated as the average of the positive scores of all the synsets of that word present in SentiWordNet. The negative score is calculated in similar fashion. Those words which are not present in

[4] All the possible POS forms present in SentiWordNet (Noun, Adjective, Verb, and Adverb) and their various senses.

SentiWordNet are assigned zero for both positive and negative scores. The objective score for each word is calculated as:

$$objScore(w_i) = 1 - (posScore(w_i) + negScore(w_i)) \qquad (1)$$

Where $posScore(w_i)$, $negScore(w_i)$, $objScore(w_i)$ are the positive, negative, objective score respectively of i^{th} word.

Step 3: Sentence Scoring

The sentence score is calculated by averaging the score of the words present in the sentence:

$$senPosScore(S) = \frac{1}{n}\sum_{i=1}^{n} posScore(w_i) \qquad (2)$$

$$senNegScore(S) = \frac{1}{n}\sum_{i=1}^{n} negScore(w_i) \qquad (3)$$

$$senObjScore(S) = \frac{1}{n}\sum_{i=1}^{n} objScore(w_i) \qquad (4)$$

Where,

- *senPosScore(S)*, *senNegScore(S)*, *senObjScore(S)* are the positive, negative, objective score respectively of sentence S.
- $posScore(w_i)$, $negScore(w_i)$, $objScore(w_i)$ are the positive, negative, objective score respectively of the i^{th} word in sentence *S*.
- n = Total no. of words in the sentence.

Step 4: Repair Heuristics for Sentence Scoring

Following repair heuristics are used to modify the sentence scores:

1. If a negation word appears in the sentence then the polarity of the word following the negation word is reversed (positive score will become negative score and vice versa). The sentence score is recalculated.
2. If intensifiers like very, more, etc. appears in the sentence then we increase the larger score of the word by a small amount ‘x’ and recalculate the sentence score. The value of ‘x’ is obtained empirically.

Step 5: Document Scoring

Positive, negative and objective scores for whole document are calculated as the average score of all the sentences:

$$docPosScore(D) = \frac{1}{n}\sum_{s=1}^{n} senPosScore(S) \qquad (5)$$

$$docNegScore(D) = \frac{1}{n}\sum_{s=1}^{n} senNegScore(S) \tag{6}$$

$$docObjScore(D) = \frac{1}{n}\sum_{s=1}^{n} senObjScore(S) \tag{7}$$

Where,

- *docPosScore(D)*, *docNegScore(D)*, *docObjScore(D)* are the positive, negative, objective score respectively of whole D^{th} document.
- *senPosScore(S)*, *senNegScore(S)*, *senObjScore(S)* are the positive, negative, objective score respectively of S^{th} sentence.
- n = Total No. of Sentences in the document.

Step 6: Rating Generation

In order to generate 5 star rating of the document we first find the maximum positive and negative scores in each approach, which are the highest and lowest ranges to generate the ratings. Then we divide this range in 10 sub ranges and calculate the normalized rating in the range 0 to 1 in step size of 0.1. In Normalized rating system 0.0 indicates 'Highly Negative' and 1.0 indicates 'Highly Positive'. Similarly in Star rating system '0 Star' and '4 Star' shows the same meaning. The 5 star rating is generated using normalized rating as follows:

$$0\ Star\ if,\ Normalized\ Rating \le 0.2 \tag{8}$$

$$1\ Star\ if,\ 0.3 \le Normalized\ Rating \le 0.4 \tag{9}$$

$$2\ Star\ if,\ 0.5 \le Normalized\ Rating \le 0.6 \tag{10}$$

$$3\ Star\ if,\ 0.7 \le Normalized\ Rating \le 0.8 \tag{11}$$

$$4\ Star\ if,\ 0.9 \le Normalized\ Rating \le 1.0 \tag{12}$$

3.2 POS based Approach

This approach is similar to the previous approach except that we utilize POS (Parts of Speech) information to assign scores to words. Instead of using the average score of all the synsets of term, the positive score and negative score for each POS category are calculated separately. The document is first tagged using Stanford POS Tagger[5] and score is calculated for each category, i.e., good#a (adjective), good#n (noun), good#r (adverb) are treated as separate words and separate scores will be calculated. SentiWordNet contains multiple synsets of terms as in WordNet 2.0 due to possible usages of that term as different POS (Parts of Speech) and multiple senses for each POS. In POS based approach the positive score and negative score of each term is

[5] Available at, http://nlp.stanford.edu/software/stanford-postagger-full-2008-09-28.tar.gz.

calculated as the score of the synset of sense–1 of any POS category[6]. For verb tokens stemming is used to calculate the positive and negative scores from SentiWordNet.

3.3. Context Sensitive Approach

This approach uses context sensitive information to modify scores calculated using “POS based approach”. The modified sentence scores are used to calculate document score. The heuristics being used in this work are discussed below:

1. Intra-sentence conjunction rule

1.1 Some conjunctions such as *“and, not only - but also, etc”* joins the similar types of sentences. Sentiment shown by sentences of both sides of these conjunctions should be same. For example, It is more likely to say *“This camera takes great pictures and has a long battery life.”* than to say *“This camera takes great pictures and has a short battery life.”*

1.2 Conjunctions such as *“but, yet, or, etc”* usually join sentences of opposite polarity. e.g., *“This camera takes great pictures but takes long time to focus.”*

If any sentence containing such type of conjunction and in one part of the sentence there is a word like *“long, short, etc”* which does not show proper orientation then the scores of this context sensitive word are related with the scores of corresponding sentiment bearing word (*great*) in another part of the sentence. Thus for the conjunctions which joins the similar types of sentences the positive and negative scores of this context sensitive word are calculated from the positive and negative scores of corresponding sentiment bearing word. While for the conjunctions which join the opposite types of sentences the positive score of this context sensitive word is calculated from the negative score of corresponding sentiment bearing word and similarly negative score is calculated.

2. Intra sentence comma rule

2.1 Sentiment of sentences joined by *comma or semicolon* is same. For example, *“The camera has a long battery life, which is great.”*

If any sentence containing comma and in one part of the sentence there is a word like *“long, short, etc”* which does not show proper orientation then the positive and negative scores of this context sensitive word are calculated from the positive and negative scores of corresponding sentiment bearing word (*great*) in another part of the sentence.

3. Inter sentence similarity rule

3.1 People usually express similar opinion across sentences, unless there is an indication of opinion change using words such as *but* and *however*. For example, *“The picture quality is amazing. The battery life is long.”* seems more natural than *“The picture quality is amazing. The battery life is short.”*

If any sentence contains context sensitive word then the corresponding sentiment bearing word is searched in prior and next sentences. The positive and negative scores

[6] In WordNet 2.0 the synset of sense–1 has highest term frequency and it is the most appropriate sense for that term, so the assumption is that using the scores of sense–1 will give the highest accuracy.

of this context sensitive word are calculated from the positive and negative scores of corresponding sentiment bearing word.

3.4 Rating Generation based on Individual Features

In this problem the rating from the review documents is generated based on some features of the product, movie or service for which the review is written. For example the movie having overall rating 2 star can be rated according to its individual features: 3 star according to story or plot, 4 star according to music, 1 star according to cast, 3 star according to cinematography etc. Thus for music lovers this movie is very good (4 star) in spite of its overall rating i.e. 2 star.

3.4.1 Selection of Features

The feature on which rating is to be generated is provided by the user. The user can input single term representing a feature or a set of keywords. For example, the user can provide the feature “story” with additional keywords like “plot”, “script”, etc. The system extracts synonyms of feature terms from WordNet 3.0[7] and offers them to user. The user may select additional terms from the list of synonyms. The manual selection of synonyms avoids inclusion of inappropriate or irrelevant synonyms. Adding synonyms increases the chances of getting feature term in reviews by making it more generalized.

3.4.2 Document Scoring according to Features

The document scores based on these features are calculated as the average scores of sentences which contain the F^{th} feature word, its related keywords, and its synonyms (if any). The Positive, negative and objective scores of document based on individual features are calculated as per the equations below:

$$featurePosScore(F_D) = \frac{1}{k}\sum_{s=1}^{k} senPosScore(S_F) \tag{13}$$

$$featureNegScore(F_D) = \frac{1}{k}\sum_{s=1}^{k} senNegScore(S_F) \tag{14}$$

$$featureObjScore(F_D) = \frac{1}{k}\sum_{s=1}^{k} senObjScore(S_F) \tag{15}$$

Where,

- *featurePosScore*(F_D), *featureNegScore*(F_D), *featureObjScore*(F_D) are the positive, negative, objective score respectively of document according to F^{th} feature in D^{th} document.
- *senPosScore*(S_F), *senNegScore*(S_F), *senObjScore*(S_F) are the positive, negative, objective score respectively of S^{th} sentence which contains feature F .
- k = Total No. of Sentences in the document which contains the F^{th} feature word.

[7] WordNet 3.0, http://wordnetcode.princeton.edu/3.0/WordNet-3.0.tar.gz.

4. The Experiments and Results

4.1 Dataset

We evaluate the proposed algorithms for rating generation on sentiment scale dataset[8]. This dataset was developed and used by Pang and Lee[18] and contains four sets of reviews each extracted from the reviews written by a single author. These reviews are put in four different directories containing 1027 reviews, 1307 reviews, 902 reviews and 1770 reviews respectively. Each directory contains subjective reviews (subj.author), the source html file name of the review from which the extract was created (id.author), 3 class ratings '0, 1, 2' (label.3class.author), 4 class ratings '0, 1, 2, 3' (label.4class.author) and normalized ratings ranging from '0.0 to 1.0' (rating.author). The reviews are in the form of paragraph and the rating of each paragraph is given in the corresponding rating file.

For the evaluation of rating generation based on individual features, there is no standard dataset available. So we have manually prepared the dataset. For this we extract 250 reviews randomly from the set of each author reviews. Thus the totals of 1000 reviews are manually examined and the ratings are generated according to individual features. The ratings are generated for four features: story, music, cast, and cinematography. All the ratings are generated in the similar format like standard dataset described above.

4.2. The Experiment

Two experiments are performed, first for the evaluation of overall rating generation, and the second for rating generation based on individual features. In the first experiment three test runs are performed. In first run the accuracy of baseline method is evaluated. The baseline in this work is SentiWordNet average scoring approach which assigns the average score of all the senses and parts of speech to a subjective term listed in SentiWordNet. The baseline method also uses two repair heuristics to handle intensifiers and negation words in the context. The second run is performed to evaluate the effect of using part of speech tagging in scoring. In this approach we tag the document before consulting SentiWordNet and use the sense–1 score of the same category only. The third run evaluates the performance of context sensitive heuristics.

In second experiment the accuracy of rating generation based on individual features is evaluated only for the context sensitive approach. Here also two test runs are performed. First run evaluates the accuracy of rating generation based on individual features without taking synonyms of features from WordNet. While in the second run the accuracy is evaluated with the consideration of synonyms of feature terms from WordNet. For each test run average accuracy for all authors is calculated. The evaluation is done in terms of percentage accuracy over all the authors.

4.3. Evaluation Measure

For evaluation, we calculate an accuracy value over the normalized rating. First, the accuracy of each review for each author is calculated and then the average accuracy of

[8] Sentiment Scale dataset, http://www.cs.cornell.edu/people/pabo/movie-review-data/scale_data.tar.gz.

each author is calculated. Finally the overall accuracy is calculated as the average accuracy over all the authors. The accuracy of each review is calculated using the following expression:

$$Accuracy = \left(1 - \left|Normalised\ Rating\ by\ our\ system - actual\ Normalized\ Rating\right|\right) * 100$$

An accuracy of 100% will be achieved when the rating assigned by our system matches with the actual rating. The accuracy calculation takes the magnitude in error into account. Misclassifying a normalized rating 0.8 as 0.2 will result in low accuracy as compared to misclassifying it as 0.5. Similarly, misclassifying a four-star rating as one star rating will result in low accuracy as compared to misclassifying it as three-star rating.

4.4 Results and Discussion

Table 1 shows the accuracies of various approaches for overall rating generation. Figure 1 compares the accuracies of these approaches.

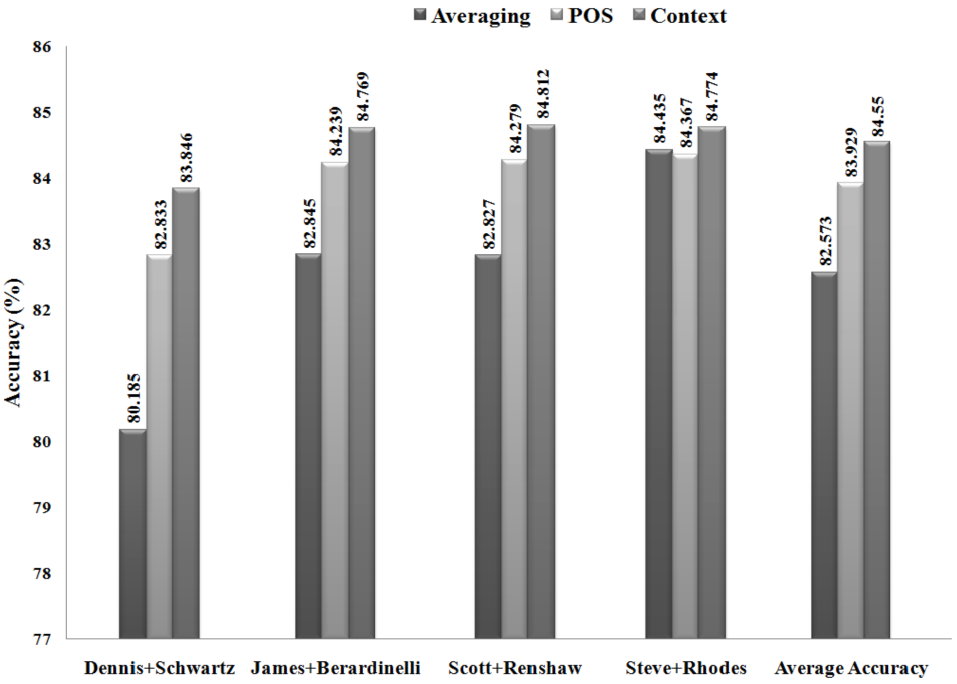

Figure 1. Comparison of Accuracy of Different Approaches for Overall Rating Generation

The maximum accuracy observed is 84.81% using context sensitive approach on the reviews given by Scott+Renshaw. This method outperforms the baseline and the POS based methods with an average accuracy of 84.55%. Pang and Lee [18] reported a maximum accuracy of 76% on the same dataset.

Table 1. Evaluation Results for Overall rating Generation

Dataset	SentiWordNet average scoring approach	POS based approach	Context Sensitive Approach
Dennis+Schwartz	80.185	82.833	83.846
James+Berardinelli	82.845	84.239	84.769
Scott+Renshaw	82.827	84.279	84.812
Steve+Rhodes	84.435	84.367	84.774
Average Accuracy	82.573	83.929	84.550

As shown in Table 1 the POS based approach performs better than SentiWordNet average scoring approach for all except one set of reviews (subj.Steve+Rhodes). For this set of reviews, SentiWordNet average scoring approach gives better accuracy. The use of part of speech tag helps in choosing the correct part of speech in the scoring process. This accounts for improved accuracy.

Table 2. Positive and Negative scores of all synsets of "good" Present in SentiWordNet

Positive Score	Negative Score	POS	Sense
0.625	0.000	Adjective	1
0.000	0.000	Adjective	2
0.875	0.000	Adjective	3
0.625	0.250	Adjective	4
0.875	0.000	Adjective	5
0.625	0.000	Adjective	6
0.875	0.000	Adjective	7
1.000	0.000	Adjective	8
0.750	0.000	Adjective	9
0.000	0.000	Adjective	10
0.375	0.000	Adjective	11
0.875	0.000	Adjective	12
0.625	0.000	Adjective	13
0.375	0.000	Adjective	14
0.250	0.375	Adjective	15
0.000	0.000	Adjective	16
0.875	0.000	Adjective	17
0.750	0.000	Adjective	18
0.625	0.000	Adjective	19
0.750	0.000	Adjective	20
0.875	0.000	Adjective	21
0.625	0.000	Adjective	22
0.375	0.500	Adjective	23
0.375	0.000	Adjective	24
0.500	0.000	Noun	1
0.875	0.000	Noun	2
0.750	0.000	Noun	3
0.875	0.000	Adverb	1
0.750	0.000	Adverb	2

Table 2 shows the scores of all synsets of "good" present in SentiWordNet. There is a large variation among scores. But most of the times, good describes the positive aspects of anything when used as an adjective. First sense of good as adjective in WordNet gives the same meaning. Thus the average score of all the synsets may not calculate the appropriate score in SentiWordNet average scoring approach. Hence, we use the score of the first sense in related POS category instead of average score over all the synsets in POS based approach. This provides correct score most of the times. By analyzing the WordNet we observe that the frequency of term categorized as sense−1 is highest as comparable to other senses in any POS category. So the cases where the word is used as in another senses will be very less. Similar thing happens for noun and adverb category.

The context sensitive approach reports best accuracy. This is because the inter sentence and intra sentence heuristics being used in this approach helps in identifying correct sentiment in a given context. The heuristics being used are domain independent and hence can be used with other dataset as well. Consider the following sentences from test dataset:

Which makes this film dumber than dumb. It's not funny, not campy. It's just dreary and unwatchable . . . a big-budget slasher film.

Here "funny" and "campy" words cannot define the polarity of expression but the previous and later sentences gives negative orientation. By Inter sentence similarity rule, second sentence can be classified as negative and actually it is showing negative sense.

In POS based approach, we do not stem adjectives, nouns and adverbs because different forms generated from the same root words may show different degree of sentiments (e.g. happy and happiness[9], home and homeless[10], etc.) and are assigned different scores in SentiWordNet. In case of verbs stemming is needed because all forms of verbs are not listed in SentiWordNet. In case of nouns without stemming some plural words will not be detected in SentiWordNet but the subjective nouns which can be used as both singular and plural are very few and have very little impact on the performance. Hence, we have applied stemming only on the verbs.

Table 3. Evaluation results for rating generation based on individual features. Feature: Story

Dataset	Without considering synonyms of features from WordNet	Considering synonyms of features from WordNet
Dennis+Schwartz	79.362	79.667
James+Berardinelli	80.156	80.864
Scott+Renshaw	80.533	80.721
Steve+Rhodes	80.183	80.788
Average Accuracy	80.059	80.510

Table 3 shows the evaluation results for rating generation based on individual features with feature story. For "story" feature we have tested with two keywords "plot" and "script". From the list of synonyms extracted from WordNet only "tale" is selected as relevant synonym.

[9] The root word for both is happy but happy is adjective, happiness is noun, and both have different sentiment scores in SentiWordNet.

[10] The root word for both is home and home is not subjective while homeless is classified as negative word in SentiWordNet.

The evaluation results for rating generation based on individual features for music feature are present in Table 4. Accuracy for "music" feature is calculated using feature word alone i.e. no keyword is used, and no relevant synonym is extracted from WordNet.

Table 4. Evaluation Results for Rating Generation based on Individual Features Feature: Music

Dataset	Without considering synonyms of features from WordNet	Considering synonyms of features from WordNet
Dennis+Schwartz	72.233	N.A.[a]
James+Berardinelli	76.428	
Scott+Renshaw	71.834	
Steve+Rhodes	77.162	
Average Accuracy	74.414	

[a] *No relevant synonym of music found in WordNet.*

Table 5 shows the evaluation results of cast feature. "Character" keyword is used with "cast" feature, and "role" is selected as relevant synonym from the extracted list of synonyms from WordNet.

Table 5. Evaluation Results for Rating Generation based on Individual Features Feature: Cast

Dataset	Without considering synonyms of features from WordNet	Considering synonyms of features from WordNet
Dennis+Schwartz	78.425	80.126
James+Berardinelli	80.359	81.433
Scott+Renshaw	79.264	80.843
Steve+Rhodes	80.167	81.256
Average Accuracy	79.554	80.915

Table 6 shows the evaluation results for rating generation based on individual features of cinematography feature. For "cinematography" feature no keyword is used. "Motion-picture-photography" and "filming" are selected from the list of synonyms extracted from WordNet.

Table 6. Evaluation Results for Rating Generation based on Individual Features Feature: Cinematography

Dataset	Without considering synonyms of features from WordNet	Considering synonyms of features from WordNet
Dennis+Schwartz	76.286	76.559
James+Berardinelli	79.152	80.352
Scott+Renshaw	76.164	76.881
Steve+Rhodes	78.329	80.116
Average Accuracy	77.483	78.477

Considering synonyms of features and keywords from WordNet increases the accuracy as it more generalizes the feature word. The accuracies of rating generation based on individual features are less than the accuracy of overall rating generation. It is observed that not only the sentences containing feature word shows sentiment about that feature, but the other sentences also shows sentiments for that feature. For example:

The story is really terrific. It is very awesome that no one guesses who the murderer was and he killed everyone.

Here, the first sentence is directly related to feature term "story" but the second sentence is also talking about the story even it does not contain feature term. For "music" feature the accuracy is lowest as compare to other features. This is because the number of sentences in the document containing feature word "music" is very less and no synonyms are being added during expansion.

5. Conclusions

In this paper, we have proposed and evaluated a knowledge-based approach for sentiment classification of document. Most of the early work on sentiment classification focuses on polarity analysis. We focus on star ratings as it provides the degree of sentiment using which two reviews can be compared. Other works focusing on rating generation use supervised methods. Our works uses purely an unsupervised approach for this task which is a novel contribution of this paper. One more novel contribution is that we have experimented with *rating generation based on individual features* which can be very useful when the person wants to select the product, movie, service or anything else according to some particular features only.

We have used average score over all the parts of speech in first approach or the score of the first sense listed in SentiWordNet in POS based approach for assigning scores to subjective terms. However, different senses of a term may convey sentiments of varying strength and sometimes indicate opposite polarity. In SentiWordNet different numerical scores has been assigned to different senses. We have used score of first sense on the assumption that the probability of occurrence of sense–1 is highest because WordNet 2.0 lists most frequent sense as first. This means the sentiment scores of terms are not being calculated accurately. A possible solution is to identify intended sense of the word in a given context and use it for calculating scores. A number of word sense disambiguation algorithms are already in place which can be used for correct sense identification.

For *rating generation based on individual features* the accuracy is not as good as for *overall rating generation*. The reason is that we calculate the ratings based only on those sentences which are directly related to the feature word, or its related keywords and synonyms. Using anaphora and cataphora resolution techniques may overcome this problem.

References

[1] S. Agrawal and T. J. Siddiqui, "Using syntactic and Contextual Information for Sentiment Polarity Analysis", Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human (ICCIT-2009), **(2009)**, Seoul, Korea**,** pp. 620–623.

[2] M. Baroni and S. Vegnaduzzo, "Identifying Subjective Adjectives through Web-based Mutual Information", Proceedings of KONVENS-04, 7th Konferenz zur Verarbeitung Naturlicher Sprache (German Conference on Natural Language Processing), **(2004)**, Vienna, AU, pp. 17–24.

[3] E. Boiy and M. F. Moens, "A machine learning approach to sentiment analysis in multilingual Webtexts", Information Retrieval, vol. 12, no. 5, **(2009)**, pp. 526–558.

[4] K. Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis", Proceedings of the International Conference on Data Engineering (ICDE 2008), Workshop on Data Engineering for Blogs, Social Media, and Web 2.0, **(2008)**, Cancun, Mexico.

[5] A. Devitt and K. Ahmad, "Sentiment Polarity Identification in Financial News: A Cohesion-based Approach", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, **(2007)**, Prague, Czech Republic, pp. 984–991.

[6] A. Esuli and F. Sebastiani, "Determining Term Subjectivity and Term Orientation for Opinion Mining", Proceedings of EACL-06, the 11th Conference of the European Chapter of the Association for Computational Linguistics, **(2006)**, Trento, Italy, pp. 193–200.

[7] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through gloss Classification", Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, **(2005)**, Bremen, DE, pp. 617–624.

[8] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining", Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation, **(2006),** Genoa, Italy.

[9] M. Gamon, A. Aue, S. Corston-Oliver and E. Ringger, "Pulse: Mining customer opinions from free text", Proceedings of 6th International Symposium on Intelligent Data Analysis, **(2005)**, Madrid, Spain, pp. 121–132.

[10] A. Ghose, P. G. Ipeirotis and A. Sundararajan, "Opinion Mining Using Econometrics: A Case Study on Reputation Systems", Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, **(2007)**, pp. 416–423.

[11] V. Hatzivassiloglou and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives", Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, **(1997)**, Madrid, ES, pp. 174–181.

[12] J. A. Horrigan, "Online shopping. Technical report", Pew Internet & American Life Project Report, **(2008)**.

[13] J. Kamps, M. Marx, R. J. Mokken and M. D. Rijke, "Using WordNet to Measure Semantic Orientation of Adjectives", Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, vol. 4, **(2004)**, Lisbon, PT, pp. 1115–1118.

[14] A. Kennedy and D. Inkpen, "Sentiment classification of movie and product reviews using context valence shifters", Computational Intelligence, vol. 22, no. 2, **(2006)**, pp. 110–125.

[15] S. M. Kim and E. Hovy, "Determining the Sentiment of Opinions", Proceedings of COLING-04, 20th International Conference on Computational Linguistics, **(2004)**, Geneva, pp. 1367–1373.

[16] B. Pang and L. Lee, "A Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", Proceedings of the ACL-2004, **(2004)**, Madrid, Spain, pp. 271–278.

[17] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, **(2008)**, pp. 1–135.

[18] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, **(2005)**.

[19] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), **(2002)**, pp. 79–86.

[20] E. Riloff, J. Wiebe and T. Wilson, "Learning Subjective Nouns using Extraction Pattern Bootstrapping", Proceedings of CONLL-03, 7th Conference on Natural Language Learning, **(2003)**, Edmonton, CA, pp. 25–32.

[21] K. Sarvabhotla, P. Pingali and V. Varma, "Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents", Information Retrieval, vol. 14, no. 3, **(2011)**, pp. 337–353.

[22] P. D. Turney, :Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, **(2002)**, Philadelphia, US, pp. 417–424.

[23] P. D. Turney and M. L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association", ACM Transactions on Information Systems, vol. 21, no. 4, **(2003)**, pp. 315–346.

[24] T. Wilson, J. Wiebe and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses", Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence, **(2004)**, San Jose, US, pp. 761–769.

Authors

Shaishav Agrawal

Shaishav Agrawal was born at Bareilly, India in 1985. He obtained Bachelors of Technology in computer science from Uttar Pradesh Technical University, Lucknow, India (2007). He completed his Masters in Technology from IIIT Allahabad, India (2009). He was Assistant Professor in Lovely Professional University, Jalandhar from 2009 to 2010. Presently he is a doctoral student in IIIT Allahabad, since September 2010. His areas of expertise are Information Retrieval, Natural Language Processing, Image Processing, Speech Recognition etc.

Tanveer J Siddiqui

Tanveer J Siddiqui obtained her PhD degree from the University of Allahabad. She has more than 12 years of experience in teaching and research. Her research area includes Natural Language Processing and Information Retrieval. She has co-authored/edited four books in the broad area of her research and published more than 25 papers in Journals and Conference proceedings.