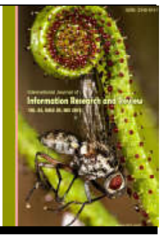
IJIRR

International Journal of Information Research and Review
Vol. 03, Issue, 05, pp. 2404-2408, May, 2016



Research Article

OPINION FEATURE SELECTION AND CLASSIFICATION FOR REDUNDANCY MINIMIZATION

***Dhivya, P., John Basha M. and Selvalakshmi, C.**

Department of CSE, PTR Engineering College, Madurai, India

ARTICLE INFO

Article History:

Received 22nd February 2016
Received in revised form
14th March 2016
Accepted 17th April 2016
Published online 30th May 2016

Keywords:

Data mining,
Feature selection,
Feature ranking,
Redundancy minimization

ABSTRACT

Feature selection, also known as attribute selection or variable subset selection, is the process of selecting a subset of relevant features. It has been the focus of interest for quite some time and much work has been done. With the creation of huge databases and the consequent requirements for good machine learning techniques, new problems arise and novel approaches to feature selection are in demand. In previous research recognized this important issue and propose, wordnet and pos tagger tools are used to minimize the redundancy between sequentially selected features by calculating efficiency and opinion strength of features. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. Our new model used to identify the polarity of features, need to extract the sentiment features from text. We are using senti wordnet tool to performing the sentiment analysis. To separate the sentiment features, we are using part of speech tagger.

INTRODUCTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. To choose a subset of input variables by eliminating features with little or no predictive information is the main idea of feature selection. Feature selection can significantly improve the comprehensibility of the resulting classifier models. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery necessary for treatment or not. Feature selection in supervised learning is a well studied one, where the main goal is to find a feature subset that produces higher classification accuracy. Recently, several researches (Dy and Brodley, 2000, Devaney and Ram, 1997, Agrawal *et al.,* 1998) have studied feature selection and clustering together with a single or unified criterion. For feature selection in unsupervised learning, to find natural grouping of the examples in the feature space the learning algorithms are designed. Thus feature selection in unsupervised learning aims to find a good subset of features that forms high quality of clusters for a given number of clusters.The majority of real-world classification problems (Kohavi and Sommerfield, 1995; Koller and Sahami, 1996) require supervised learning where the underlying class probabilities and class-conditional probabilities are unknown, and each instance is associated with a class label (Dash and Liu, 1997). In real-world, we often have knowledge about relevant features and irrelevant features. Many types of features are introduced the domain and conclusion in the subsistence of irrelevant/redundant features to the decide on concept. A relevant feature is neither irrelevant nor redundant to the decide on concept; an irrelevant feature is indirectly associate with the decide on concept, not directly but affect the learning process, and a irrelevant and redundant feature does not add anything new to decide on concept (Zakaria Elberrichi *et al.,* 2008). In many selection and classification problems, it is difficult to understand good classifiers before removing these irrelevant features due to the large number of the data. Reducing the number of irrelevant/redundant features can reduce the computation time of the learning phase and permit a more general classifier. This helps in getting a better awareness into the underlying concept of a real-world classification problem.

Related works

A large number of researchers had proposed opinion analysis for feature selection and classification for redundancy minimization. In previous research, Ding et al. recognized this particular issue and proposed the mRMR (minimum Redundancy Maximum Relevance Feature Selection) model to minimize the redundancy among the sequentially selected features. However, this method used the greedy search, where the global feature redundancy wasn't considered and the results

****Corresponding author: Dhivya, P.,***
Department of CSE, PTR Engineering college, Madurai, India.

are not optimal. Zakaria Elberrichi, Abdelattif Rahmoun, and Mohamed Amine Bentaalah (Muhammad Zubair Asghar *et al.,* 2014) paper explores a method to categorize the text documents that use WordNet concept. The bag of words representation used for text representation is unsatisfactory as it ignores possible relations between terms. The proposed method extracts generic concepts from WordNet for all the terms in the text and then it forms a new representative vector by combining them with the term in different ways. In (Hu and Lium 2004), the author says feature based sentiment analysis include feature extraction**,** sentiment prediction, sentiment classification and optional summarization modules (Amitava Das *et al.,* 2008). Feature extraction process generates the extracted features by taking text as input and generate it in any of the forms like Lexico-Syntactic or Stylistic, Syntactic and Discourse based (Ahmed Abbasi *et al.,* 2006; Hu and Liu, 2004; Turney, 2002) paper focuses on online customer reviews of products. It makes two contributions. First, it proposes a novel framework for comparing and analyzing consumer opinions of competing products. A prototype system called Opinion Observer is also implemented. The system is such that with a single glance of its visualization, the user is able to clearly see the strengths and weaknesses of each product in the minds of consumers in terms of various product features. This comparison is useful to both product manufacturers and potential customers. Peter D.Turney (Wang *et al.,* 2014) proposed with the explosive growth of the social media content on the Internet in the past few years, people now express their views on almost anything in discussion. Finding and monitoring the opinion sites on the web is a difficult task. Thus there is a need for automatic opinion summarization systems and discovery. Sentiment Analysis or Opinion Mining is the computational study of opinions, sentiments and emotions expressed in text. This paper describes the field of Sentiment Analysis and its latest developments. However, finding opinion sites and monitoring the Web can still be a formidable task because there are a large number of diverse sites, and each site may also have a huge volume of opinionated text. In many cases, opinions are hidden in blogs and long forum posts. It is difficult for a human reader to find relevant sites, extract related sentences with opinions, read, summarize and organize them into usable forms. Thus the Automated opinion discovery and summarization systems are needed. F.Y. Wang (Lloret *et al.,* 2010) addresses the difficulties to associate with their development, operation. In (Samaneh Moghaddam, ?), the authors utilize A comparative opinion expresses a preference relation of two or more objects based on some of their shared features.

Pos Tagging

We introduce sentiment analysis as a pre-classification procedure to support subjectivity-aware POS tagging. The POS corpus is classified into two categories as “subjective” or “objective”. This pre-classification is typically used for sentiment analysis. POS tagging can benefited by subjectivity classification. The genre information can also help POS tagging, but found out that genre information in brown corpus is not as beneficial for POS tagging as subjectivity information. Besides for a collection of article for each kind of genres it’s difficult to collect. We use two collections of subjectivity to build the upper and lower bound of (e A; e B; e_). We found out that the subjectivity labeling is easier than POS tagging. The approach is to contribute more when labeled with “subjective” or “objective” tag. By supervised learning model can be trained using the pre-classified data. In (e *A;* e *B;* e_), the lower bound is calculated from training set with “objective” pre-class, and the upper bound is from the “subjective” pre-class.

A. Datasets

The data set we use is based on review POS which is a part-of-speech corpus on 1500 reviews. The words are turned into lowercase. There are 25 tags in review POS, some of which are seldom used in common text tagging, i.e., “E” for “Emotion” such as “:)”.

B. Experimental Settings

The accuracy of the prediction of token-tag pairs is the performance metric used in this study. Since we need a corpus that have subjectivity label as well as POS tags, we tag each message as “subjective” or “objective” as the training set for Naive Bayes model. The words are only used as features to train the model. The tags that are irrelevant to subjectivity analysis are removed. The tags we used are shown in Table I. There are more interjection and adjectives in subjective reviews.

Table 1. The User reviews part-of-speech tags that used in subjectivity Analysis

Tag	Description	Example	% in objective	% in subjective
A	adjective	great	6.5	4.5
R	adverb	very	5.2	4.5
!	interjection	lol	4.6	1.7
E	emotion	:-)	1.2	0.8
,	punctuation	!!!	12.7	11.7
G	abbreviation	ily	0.6	1.1
V	verbal	want	14.9	15.0
N	common noun	gift	12.5	14.3

For POS tagging task at semantic level the sentimental information provides some global knowledge. The experiments illustrate that the subjectivity classification can benefit the POS tagging task. In an attempt to use existing or calibrated global information rather than to generate detailed features the proposed approach improves the POS tagging performance. Interval-type HMM allows un-determined model parameters to cope with global information.

Lexical database

A lexical database is a lexical resource that permits access to its contents which has an associated software environment database. The database may be a general-purpose database into which lexical information has been entered or a custom-designed for the lexical information. Information stored in a database includes lexical category and synonyms of words, as well as phonological and semantic relations between different words or sets of words.

A. Wordnet

Word Net is a thesaurus for the English language developed at the University of Princeton and based on psycholinguistics studies. It was originated as a data-processing resource which covers synsets. The synsets are sets of synonyms which gather lexical items having similar significances. The definition of the synsets varies from the specific one to the very general. The most general synsets cover a very broad number of significances whereas the most specific synsets gather a restricted number of lexical significances. The difference which has WordNet compared to the traditional dictionaries is the separation of the data into four data bases associated with the categories of verbs, nouns, adjectives and adverbs. Each database is differently organized than the others. The names are orderly in hierarchy, the verbs by relations, the adjectives and the adverbs by N-dimension hyperspaces. The approach suggested is composed of two stages. The first stage relates to the learning phase and it consists of generating a new text representation based on merging terms with their associated concept and also selecting the characteristic features for creating the categories profiles. The second stage relates to the classification phase which consists of weighting the features in the categories profiles and calculating the distance between the categories profiles and the profile of the document to be classified.

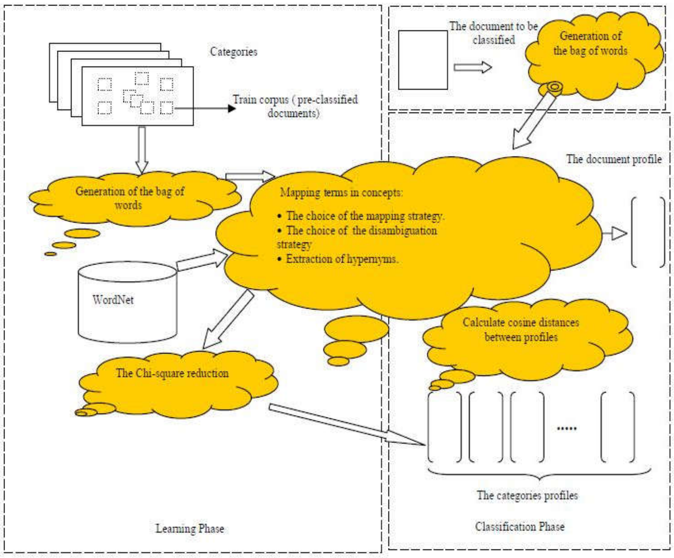

Categories
Train corpus (pre-classified documents)
Generation of the bag of words
Mapping terms in concepts:
• The choice of the mapping strategy.
• The choice of the disambiguation strategy
• Extraction of hypernyms.
WordNet
The Chi-square reduction
Learning Phase
The document to be classified
Generation of the bag of words
The document profile
Calculate cosine distances between profiles
.....
The categories profiles
Classification Phase

Figure 1. The suggested approach.

MATERIALS AND METHODS

In the proposed methodology, we present our pre-classification model training algorithm, which consists of a preprocessing phase and a mining phase. In the preprocessing phase, the relevant dataset is extracted from the original database. In the mining phase, given dataset is performed with POS tagging and extraction of sentiment features.

Opinion analysis using wordnet tool is performed on the extracted sentiment features. During the opinion analysis, opinion strength of sentiment features is calculated by use of wordnet tool. Senti Word Net assigns to each synset of Word Net three sentiment scores. Based on opinion strength of sentiment features we have classifying the word into five classes. 1. Strong positive, 2. Weak positive 3. Strong negative 4. Weak negative 5. Neutral.

A. Pseudo code

This consists of two phases.

Step 1: Preprocessing phase

Extract some statistical information from the original database

Step 2: Mining phase

We apply POS tagging in the extracted statistical information. The pre-classification model training algorithm is stated in Algorithm 1. Lines 1-6 prepare the review corpus. The initial subjectivity labels are crafted by hand and then the POS review corpus can be expressed as (*SSobj ; SSsub*).

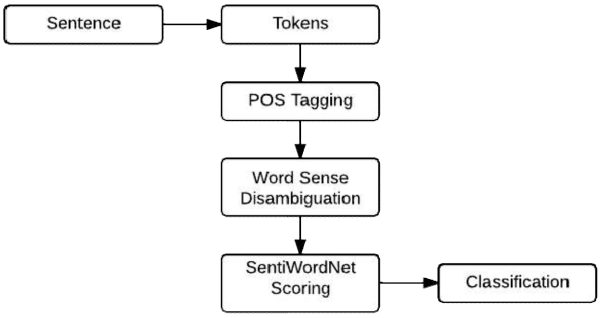


Figure 2. The Proposed method Diagram

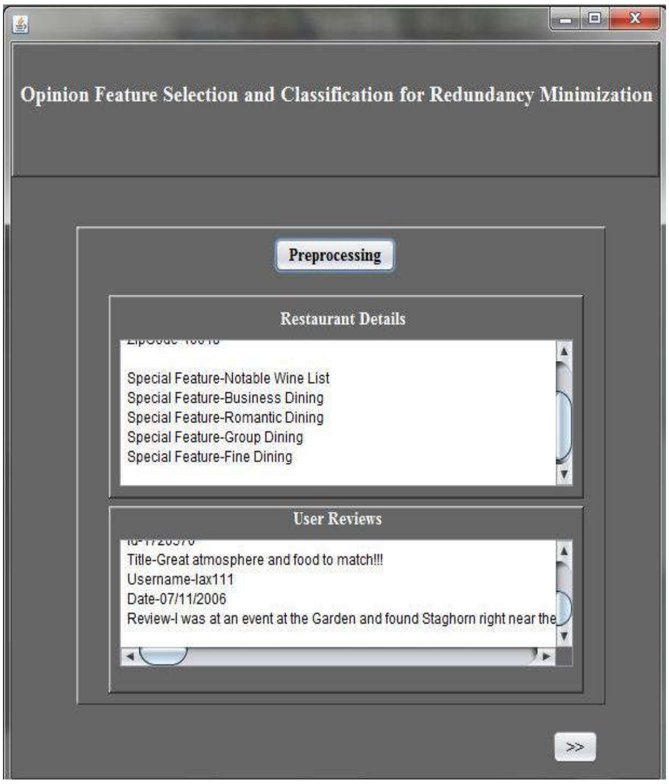
By turning words into lower cases and replacing special names the corpus is cleaned. A line 7-12 generates a feature set using word features and with the training set trains the model. The NB parameters are typically trained according to Equs. (3), (4) and (5). In Lines 13-20, (*SSobj; SSsub*) is re-classified into (*RSSobj; RSSsub*) using the NB model to get a pre-classified training set

Algorithm 1 Pre-classification model training

```
Input: POS-tagging corpus for Restaurant Reviews
Output: Find Most Effective Features,
Pre-classified corpus to train Minimization Redundancy
01. Prepare Review corpus:
02. Initial subjectivity labeling
03. by hand: $Category = 'obj';' sub'$
04. $SS_{\_obj} = [(Word,\ Tag)^{T}\,]^{M}$
05. $SS_{\_sub} = [(Word,\ Tag)^{T}\,]^{M}$
06. Clean the corpus
07. Model training:
08. Feature generation:
09. $feature\ (SS_{\_obj};\ SS_{\_sub,\ word\_features})$
10. $Training\ set = \{feature,\ true/false;\ obj/sub\}$
11. Evaluate the parameters using POST
12. $POST$ according to Equ.(3,4,5)
13. Pre-classification:
14. $RSS\_obj = \{\}$
15. For each $sentence$ in $\cup\{SS_{obj}\ ;\ SS_{sub}\}$
16. $if P(y = 1 \mid O) < 0.5$:
17. $RSS_{obj} = RSS_{obj} \cup \{sentence\}$
18. else:
19. $RSS_{sub} = RSS_{sub} \cup \{sentence\}$
20. Return: POST; $RSS_{obj}$; $RSS_{sub}$
```

Then the Sentiment words are extracted after applying the POS tagging.

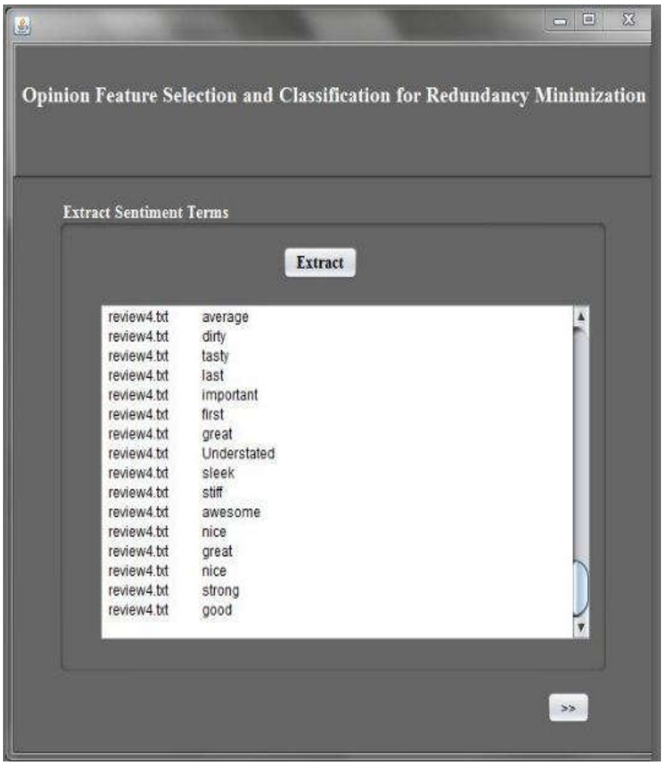
Step 3: Opinion strength calculation: The opinion strength is calculated using the wordnet tool which consists of two phases:


Opinion Feature Selection and Classification for Redundancy Minimization
Preprocessing
Restaurant Details
User Reviews
>>

```
Special Feature-Notable Wine List
Special Feature-Business Dining
Special Feature-Romantic Dining
Special Feature-Group Dining
Special Feature-Fine Dining
```

```
Title-Great atmosphere and food to match!!!
Username-lax111
Date-07/11/2006
Review-I was at an event at the Garden and found Staghorn right near the
```

Fig. 4. Pre processing


Opinion Feature Selection and Classification for Redundancy Minimization
Extract Sentiment Terms
Extract
>>

```
review4.txt     average
review4.txt     dirty
review4.txt     tasty
review4.txt     last
review4.txt     important
review4.txt     first
review4.txt     great
review4.txt     Understated
review4.txt     sleek
review4.txt     stiff
review4.txt     awesome
review4.txt     nice
review4.txt     great
review4.txt     nice
review4.txt     strong
review4.txt     good
```

Fig. 5. Sentiment terms extraction

A. Learning Phase: A text representation is generated based on merging terms with their associated concept and creating the categories profiles by selecting the characteristic features.

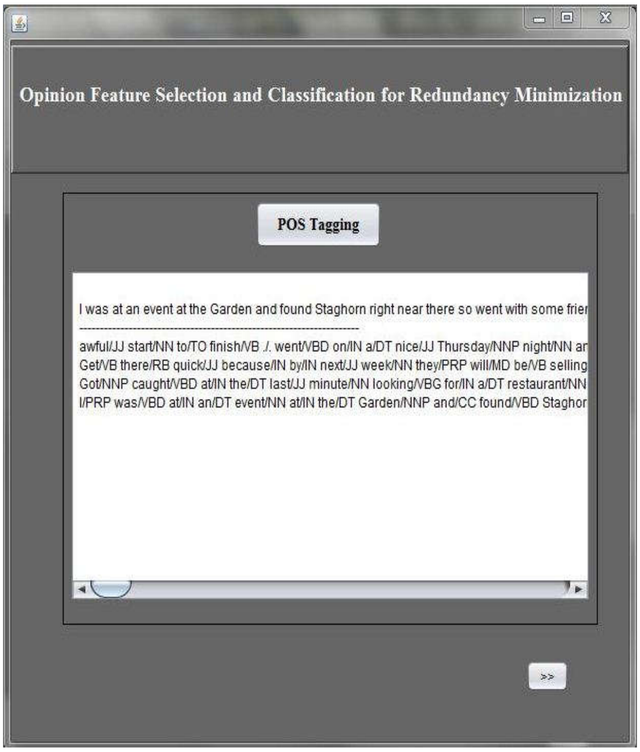
B. Classification Phase: Weighting the features in the categories profiles and calculating the opinion strength between the categories profiles and the profile of the document to be classified

Step 4: Opinion Analysis

Polarity of a word refers to its strength typically in a 'positive' vs. 'negative' sense. Here we consider three polarity levels for adjectives: positive, negative and neutral. An adjective can imply positive meaning, like 'excellent', negative meaning, like 'poor' and neutral meaning, like 'mediocre'. For training and testing the classifier (Turney, 2002) we use a set of tagged adjectives. In this set there are 30 adjectives, 10 of them are tagged as positive, 10 as negative and 10 as neutral adjectives. The input of classifier is a set of tuples containing the similarity values between the given adjective and the three fixed adjectives 'excellent', 'mediocre' and 'poor'. To predict the polarity of other adjectives we use these three fixed adjectives as reference points. The polarity is determined by aggregating the polarity of the extracted adjectives based on their frequencies.

Table 2. Tagged adjectives used for training and testing the classifier

Positive	Neutral	Negative
good	mediocre	bad
nice	average	terrible
awesome	enough	weak
excellent	fair	bitter
great	okay	imperfect
precious	fine	poor
satisfactory	neutral	faulty
exceptional	ordinary	defective
outstanding	reasonable	awful


Opinion Feature Selection and Classification for Redundancy Minimization
POS Tagging
>>

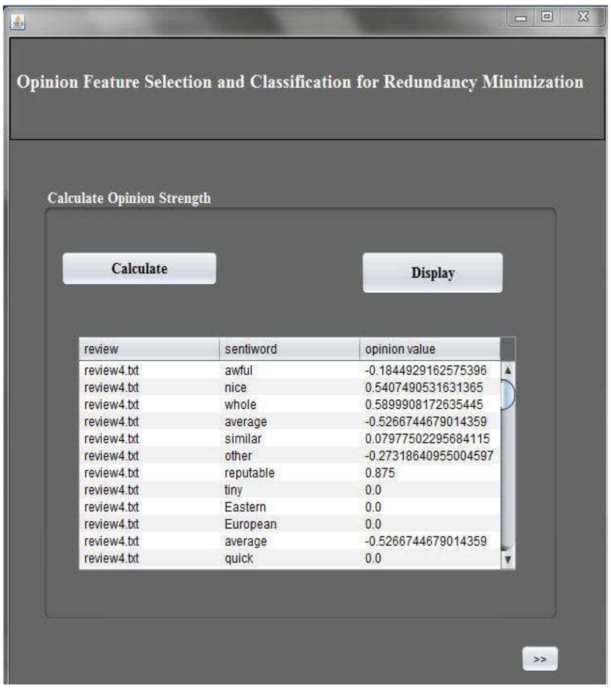
```
I was at an event at the Garden and found Staghorn right near there so went with some frier
-----------------------------------------------------------------
awful/JJ start/NN to/TO finish/VB ./. went/VBD on/IN a/DT nice/JJ Thursday/NNP night/NN ar
Get/VB there/RB quick/JJ because/IN by/IN next/JJ week/NN they/PRP will/MD be/VB selling
Got/NNP caught/VBD at/IN the/DT last/JJ minute/NN looking/VBG for/IN a/DT restaurant/NN
I/PRP was/VBD at/IN an/DT event/NN at/IN the/DT Garden/NNP and/CC found/VBD Staghor
```

Fig. 6. POS Tagging

Each adjective is also assigned a weight which is equal to its frequency in that review. The weighted average of the adjective scores is computed to find the polarity of the review is then determined. A specific classifier is needed for each context. When the method learns the opinion classifier, it can be applied to other contexts.

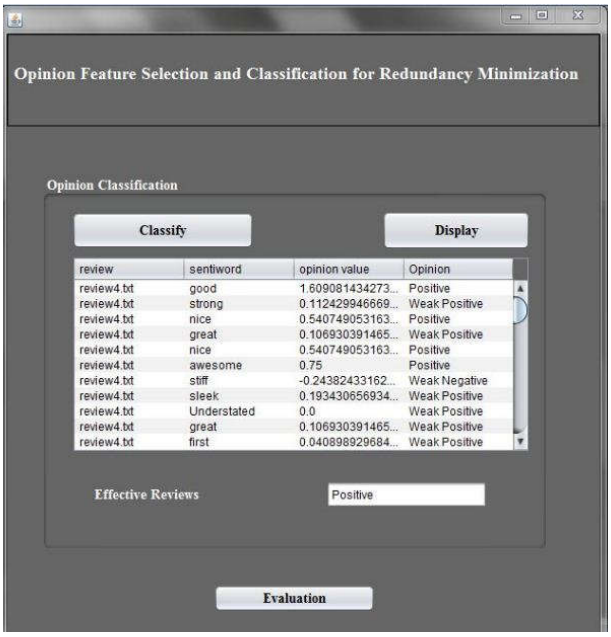
EXPERIMENTAL RESULTS

In these experimental results, we collect sample data from the restaurant user review set and performed it by proposed methods. The below figures show the proposed method result.


Opinion Feature Selection and Classification for Redundancy Minimization
Calculate Opinion Strength
Calculate
Display
>>

review	sentiword	opinion value
review4.txt	awful	-0.1844929162575396
review4.txt	nice	0.5407490531631365
review4.txt	whole	0.5899908172635445
review4.txt	average	-0.5266744679014359
review4.txt	similar	0.07977502295684115
review4.txt	other	-0.27318640955004597
review4.txt	reputable	0.875
review4.txt	tiny	0.0
review4.txt	Eastern	0.0
review4.txt	European	0.0
review4.txt	average	-0.5266744679014359
review4.txt	quick	0.0

Fig. 7. Opinion strength calculation


Opinion Feature Selection and Classification for Redundancy Minimization
Opinion Classification
Classify
Display
Evaluation

review	sentiword	opinion value	Opinion
review4.txt	good	1.609081434273...	Positive
review4.txt	strong	0.112429946669...	Weak Positive
review4.txt	nice	0.540749053163...	Positive
review4.txt	great	0.106930391465...	Weak Positive
review4.txt	nice	0.540749053163...	Positive
review4.txt	awesome	0.75	Positive
review4.txt	stiff	-0.24382433162...	Weak Negative
review4.txt	sleek	0.193430656934...	Weak Positive
review4.txt	Understated	0.0	Weak Positive
review4.txt	great	0.106930391465...	Weak Positive
review4.txt	first	0.040898929684...	Weak Positive

Effective Reviews: Positive

Fig. 8. Opinion Analysis process

In this process, we have to select the review data for processing. Based on opinion mining we find the effective review. Classification rules are generated based for review data set. Sentiment analysis used to find out the polarity of features.

Conclusion

This paper introduces a technique for classifying a review as positive, negative, or neutral. The core of the technique is the phase where we uses WordNet to compute the similarity values between two adjectives, and then the values are used to learn the classifier for predicting polarity of each adjective. In experiments with 100 reviews from the movie reviews corpus in NLTK, our algorithm attains an accuracy of 73% while the baseline method in the best case can attain an accuracy of 64%. It can attain high accuracy using a small training set is the key advantage of our method. In addition, our proposed opinion polarity classifier is independent from the context and can be applied to different review types. On the other hand, the baseline classifier totally depends on the context which limits its usage. In the end, we can say that high accuracy along with simplicity of our method may encourage further work with opinion polarity.

REFERENCES

Ahmed Abbasi, etl. “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums” ACM Transactions on Information Systems, Vol. 26, No. 3, Article 12, 2008.

Amitava Das et al, 2008, Topic-Based Bengali Opinion Summarization, Coling 2008: Poster Volume, pages 232–240, Beijing, August 2010.

Hu, M., and Liu, B. 2004. Mining Opinion Features in Customer Reviews. *AAAI’04*, 2004.

Hu, M., and Liu, B. 2004. Mining Opinion Features in Customer Reviews. *AAAI’04*, 2004.

Kohavi, R. and Sommerfield, D., Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In: Proceedings of First International Conference on Knowledge Discovery and Data Mining, Morgan Kaufmann, 192–197, 1995.

Koller, D. and Sahami, M., Toward optimal feature selection. In: Proceedings of International Conference on MachineLearning, 1996.

Lloret, E., Saggion, H. & Palomar, M. Experiments on summary-based opinion classification, *in* ‘Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysisand Generation of Emotion in Text’, p. 107-115.

M. Dash and H. Liu. Feature selection for classification. Intelligent data analysis, 1(1-4):131–156, 1997.

Muhammad Zubair Asghar1, Aurangzeb Khan2, Shakeel Ahmad1, Fazal Masud Kundi1 “A Review of Feature Extraction in Sentiment Analysis,” in Proc, 2014, pp. 181–183.

Samaneh Moghaddam, Fred Popowich Experiments on summary-based opinion Opinion Polarity Identification through Adjectives

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 417-424

Wang, D., F. Nie, and H. Huang. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In Machine Learning and Knowledge Discovery in Databases, pages 306–321. Springer, 2014.

Zakaria Elberrichi, Abdelattif Rahmoun, and Mohamed Amine Bentaalah 2008. Knowledge Discovery in Databases: Using WordNet for Text Categorization 2008, pp.16-17.

****