Identification of Opinionated Text and Classifying its Sentiments of Online Product Reviews Using Corpus-Based Approach: Opinion Mining

Regina. J
Dept.of CSE, BNMIT, Bangalore
E-mail : john_regi@rediffmail.com

Abstract – **Opinion miming aims to determine the opinion of the author with respect to some topic or the overall contextual polarity of a document by classifying the sentiment as positive, negative or neutral. It can help researchers to study opinion and sentiment information on the internet by identifying and analyzing texts containing opinion and emotions. This method identifies the opinionated texts as subjective or objective and classifies the subjective text as positive, negative and neutral. The proposed method adopted corpus-based approach to extract opinion word list. This method combines both machine learning and sentiment orientation approaches. The proposed method also includes nouns as the part-of-speech (pos) which is more context-dependent in addition to adverbs, adjectives and verbs and also it includes content-free, content-specific and sentiment features to improve the sentiment classification performance. The proposed method has three main tasks: Data acquisition, Feature generation and Classification and Evaluation. The performance of this method is verified on online product review articles.**

Keywords – Sentiment classification, feature representation, online product reviews, content-free, content-specific and sentiment features.

I. INTRODUCTION

As an emerging communication platform, web 2.0 has led the internet to become more and more user centric. People are participating in and exchanging opinions through online community-based social media, such as discussion boards, web forums and blogs. Along with such trends, an increasing amount of user generated content containing rich opinion and sentiment information has appeared on the internet. Understanding such opinion and sentiment information has become increasingly important for both service or product providers and users since it has played an important role in influencing consumer purchasing decisions.

A. Problem Description

Sentiment classification techniques can be used to study the opinion and sentiment information on the internet by identifying and analyzing texts containing opinions and emotions to determine whether a text is objective or subjective and whether a subjective test contains positive and negative sentiments [3,4]. The approaches the have been adopted in previous sentiment classification studies, to compile or collect the opinion word list are: corpus-based and dictionary-based [11, 22]. Dictionary-based approach typically uses WORDNET's synsets [10, 13] and hierarchies to acquire opinion words, but do not find context dependent opinion words [11]. Corpus-based approach, rely on syntactic or co- occurrence patterns in large corpora, which finds context-dependent opinion words [11, 22]. This method combines both dictionary-based and corpus-based approaches into one framework to improve sentiment classification performance [7]. To extract features from reviews, most of the studies have adopted content-free features, content-specific features and sentiment features which are referred as frequent features [9, 25, 23], which most of the people talked about. Few studies have shown there are some features, which only a small number of people talked about is referred as an infrequent features [22]. This feature can also be interesting to some potential customers [7]. The proposed method, incorporate both frequent features and infrequent features to improve sentiment classification performance.

II. RELATED WORK

A. *Sentiment classification approaches*

In general, sentiment analysis is concerned with analysis of direction-based text, i.e. text containing opinions and emotions [25]. Sentiment classification studies attempt to determine whether a text is objective or subjective, or whether a subjective text contains positive or negative sentiments. The common two class problem involves classifying sentiments as positive or negative [3, 4]. Additional variations include classifying sentiments as opinionated/subjective or factual/objective [8]. Some studies have attempted to classify emotions, including happiness, sadness, anger, horror etc., instead of sentiments [5]. Two approaches have been utilized in previous sentiment classification studies: machine learning [3] and semantic orientation [4, 5]. Involving text classification techniques, the machine learning approach treats the sentiment classification problem as a topic-based text classification problem [11]. Any text classification algorithm can be employed, e.g., Naïve Bayes, SVM, etc. Pang et al. [4] experimented with this approach to classify movie reviews into two classes: positive and negative. Different from the machine learning approach, the semantic orientation approach performs classification based on positive and negative sentiment words and phrases contained in each evaluation text and no prior training is required in order to mine the data [11, 8].

Two types of techniques have been used in previous semantic orientation approach based sentiment classification research, including: (1) corpus-based techniques and (2) dictionary-based techniques [11]. The corpus-based techniques aim to find co-occurrence patterns of words to determine their sentiments. Different strategies are developed to determine sentiments. For example, Turney [4] calculated a phrase’s semantic orientation to be the mutual information between the phrase and the word “excellent” (as the positive polarity) minus the mutual information between the phrase and the word “poor” (as the negative polarity). Riloff and Wiebe [5] used a bootstrapping process to learn linguistically rich patterns of subjective expressions in order to classify subjective expressions from objective expressions. Starting with a set of objective patterns adopted from previous literature, the process used a pattern extraction algorithm to learn potential subjective patterns. The learned patterns were then used to decide whether an expression was subjective or not. Dictionary-based techniques, which are another type of techniques, utilize synonyms, antonyms and hierarchies in WordNet (or other lexicons with sentiment information) to determine word sentiments [11]. Building upon WordNet, SentiWordNet [13] is a lexical resource for sentiment analysis which has more sentiment related features than WordNet. It assigns to each synsets of WordNet three sentiment scores regarding positivity, negativity, and objectivity respectively. SentiWordNet has been used as the lexicon in recent sentiment classification studies [12-14].

B. *Sentiment Classification Features*

There are mainly three categories of features that have been adopted in previous sentiment classification studies: (1) content-free features, (2) content-specific features and (3) sentiment features. Content-free features include lexical features, syntactic features, and structural features [9, 23, 25]. Lexical features are character-, or word-based statistical measures of lexical variation [23]. They mainly include: character-based lexical features, vocabulary richness measures, and word-based lexical features. Syntactic features indicate the patterns used to form sentences [23]. They mainly include: function words, punctuation, and part-of-speech. Structural features show the text organization and layout. Other structural features include technical features such as the use of various file extensions, fonts, sizes, and colors. Content-specific features are comprised of important keywords and phrases on certain topics, such as word n-grams. Previous studies have shown that content-specific features are helpful in improving text classification performance [9].In previous semantic orientation approach based sentiment classification studies, the overall sentiment of a text is determined by the sentiments of a group of words and/or phrases appearing in the text. Different categories of words or phrases have been used to determine the overall sentiment of a text. For example, Hatzivassiloglou and Wiebe [2] used different types of adjectives appearing in a text; Hu and Liu [7] also used adjectives; Turney [4] used all the two-word phrases that contained adjectives or adverbs in a given text; and Denecke [13] used the combination of adverbs, adjectives, verbs, and nouns .In this method, we incorporate them into the infrequent features [7, 11] as an additional dimension of features. However, not all features are necessary or sufficient to learn the concept of interest. Therefore, feature selection, which aims at identifying a minimal-sized subset of features relevant to the target concept, can be applied [1]. A feature selection method generates different candidates from the feature space [17] and assesses them based on some evaluation criterion to find the best feature subset.

III. MOTIVATION

Each of the two sentiment classification approaches has its advantages and disadvantages. The machine learning approach tends to be more accurate than the semantic orientation approach since a machine learning model is always tuned to the training data set, thus making it domain dependent [4,11]. If applied elsewhere, training on the new data sets is needed. In contrast, the semantic orientation approach is domain independent; no prior training is needed [11]. Therefore, it has better generality. But its classification accuracy is often not as high as that of the machine learning approach. In addition, the corpus-based techniques for semantic orientation approach often rely on a large corpus to calculate the statistical information needed to decide the sentiment orientation for each word or phrase. Corpus-based approach gives context dependent opinion words [11]. But a good lexicon is critical for the dictionary-based techniques [11]. Few studies have investigated the combination of both the machine learning approach and the semantic orientation approach to improve sentiment classification performance [25].

Motivated by the above discussion, in this study we developed a corpus-based approach, and combine both machine learning and semantic orientation approaches into one framework. Specifically, we generated a set of sentiment words based on a sentiment lexicon and used them as a new dimension of features, sentiment features, for the machine learning classifiers. Our proposed method incorporates the content-free and content-specific features used in the existing machine learning approach and also it incorporates the sentiment-features and infrequent-features in the semantic orientation approach. We demonstrated the efficacy of our proposed method by testing it on different online product review data sets.

IV. DESIGN AND IMPLEMENTATION

The corpus-based approach for sentiment classification proposed in this study consists of three main modules, Data acquisition, Feature generation and Classification and Evaluation.

A. Data Acquisition

In this module, collected datasets are parsed and stored in a database. Here, we use online product reviews as the application domain, because of reviews' increasing importance in influencing individuals' purchase decision. We carry out the experiment on the labeled product reviews from two domains: Digital cameras and mobile phones. Each domain contains equal number of positive and negative reviews from the sites: epinoins.com, dpreview.com, steves-digicam.com and amazon.com.

B. Feature Generation

In this module, three types of features are used in our proposed sentiment classification method, including content-free features (F1), content-specific (F2), sentiment features (F3). Among them F1 and F2 are from the machine learning approach and F3 are semantic oriented approach. Each test bed utilizes lexical features, syntactic features and structures features for F1 features. Unigrams and bi-grams were used as F2 features. Semantically empty stop words should be removed, the number of F2 feature varies for each text bed and is much larger than that of F1 features.

Feature Extraction and Sentiment Score Calculation

1. To parse each sentence and yield the part-of-speech (POS) tag of each word, (i.e., whether the word is a noun, verb, adjective and adverb etc.,) use Stanford POS-tagger to perform the tagging.

2. To determine the sentiment scores of the extracted adjectives, adverbs, verbs and nouns we use SENTI-WORDNET.

3. SENTI-WORDNET assigns to each synsets in WORDNET three sentiment scores, i) Positive ii) Negative and iii) Objective

4. Then, calculate the average polarity scores for each synsets separately using the prior-polarity formula,

$$\mathrm{Score}(word{=}POS)i{=}(k\text{€}SentiWordNet(word\_POS\&polarity i))\,\frac{(SentiWordNet\_Score(k)i)}{(|synsets\,(word\_POS)|)}$$

POS € {adjective, adverb, verb, noun}

i € {positive, negative, objective}

k € synsets of a given word in a particular sense

Sentiment Feature-Calculation Strategy

1. To filter out less subjective words, use midpoint of the 0-1 score scale (i.e.,0.5)

   i) If (score (word=POS) $_{\text{objective}}$ > 0.5), then consider the word as objective. Otherwise,

   ii) If ((score (word=POS) $_{\text{positive}}$) > (score (word=POS)$_{\text{negative}}$)), then consider the word in the given POS as a positive sentiment feature. Otherwise,

   iii) Consider the word in the given POS sense as a negative sentiment feature.

2. In addition, exclude the words whose positive scores are equal to the negative scores from the sentiment feature, since they do not show clear polarity tendency.

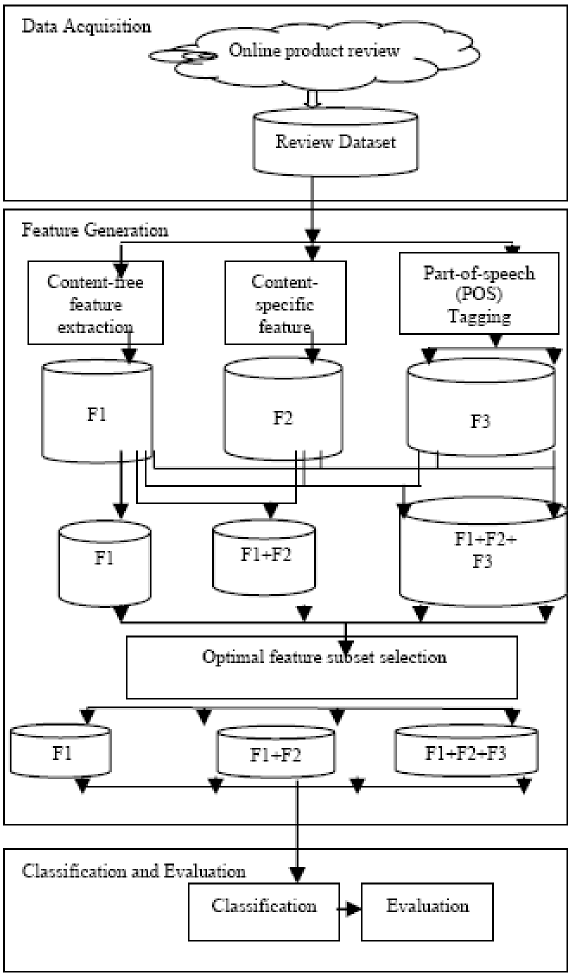

Fig.1 : Design of corpus-based approach for sentiment classification

C. Classification and Evaluation

To examine the corpus-based approach for sentiment classification, compare the performances of different feature sets using SVM as the classifier because of its reported performance in previous sentiment analysis studies. For each test bed, randomly choose 90% of reviews of training data and the remaining 10% as testing data for the train/test split.10-fold cross validation is used to evaluate. Summarize the performance measures in terms of overall accuracy, average precision, average recall and average F-measure for all the given test beds. Among the four types of features: F1, F3 and F4 are domain independent and F2 are domain dependent.

To test the performance of different types of features, we create four different feature sets in an incremental way, 1) F1 alone includes content-free features, 2) Feature set F1+F2 which consists of content-free and content-specific features, 3) F1+F2+F3 which is composed of content-free, content-specific and sentiment features . Here F1, F3 are domain independent and so the feature sets F1, F1+F3 are domain independent and the feature sets F1+F2 ,F1+F2+F3 are domain dependent. When the number of feature is large, feature selection may improve the classification performance by selecting optimal subset of features. Thus building three selected feature sets: F1, F1+F2, F1+F2+F3 to study the effectiveness of proposed sentiment classification method.

Word score calculation:

E.g,[‘best’ ,ADV] no.of synsets=3

Data_set	Positive	Negative	Objective
#1	0.5	0	0.5
#2	0	0	1
#3	0	0	1
Tot.Score	0.5	0	2.5
Word score	0.1667	0	0.833

Table:1 -Result: ”Positive” sentiment feature

Product Name	Domain	Review Category	Accuracy	Precision	Recall	F-Measure
Canon Power Shot SD1400 (cps 1400)	C1	Positive(10)	1.000000	1.000000	0.769231	0.869565
		Negative(10)	0.700000	0.700000	1.000000	0.823529
	C2	Positive(10)	1.000000	1.000000	0.758621	0.862745
		Negative(10)	0.681818	0.681818	1.000000	0.810811
Canon Power Shot D10 (cpsd 10)	C1	Positive(10)	0.900000	0.900000	0.818182	0.857143
		Negative(10)	0.800000	0.800000	0.888889	0.842505
	C2	Positive(10)	1.000000	1.000000	0.714286	0.833333
		Negative(10)	0.600000	0.600000	1.000000	0.75000

Table 2: Classification accuracy of various reviews

.RESULT: Negative –Classification bias(i.e) high precision for positive reviews

V. CONCLUSION

This method used a corpus-based approach to generate sentiment features and also included infrequent features to improve the effectiveness of sentiment classification methods. Further research, can also explore other sentiment features generation methods and compare their performance. In addition, feature selection on large feature sets can be shown to improve the classification performance on relatively large data sets. Text classification could be an additional future research direction. Moreover, although we used English language review data in this study, the proposed method can also be applied to other languages, and a multilingual sentiment-based lexicon needs to be developed in the future.

VI. REFERENCES

[1] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis: An Int'l J., vol. 1, no. 3, 1997,pp. 131–156.

[2] V. Hatzivassiloglou and J. Wiebe, "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," Proc. 18th Int'l Conf. Computational Linguistics, ACL Press, 2000, pp. 299–305.

[3] B. Pang et al., "Thumbs up? Sentiment Classification Using Machine Learning Techniques," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), ACL Press, 2002, pp. 79–86.

[4] P.D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,"Proc. 40th Ann. Meetings of the Assoc. Computational Linguistics, ACL Press, 2002, pp. 417–424.

[5] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," Proc. Conf. Empirical Methods in Natural Language Processing, ACL Press, 2003, pp. 105–112.

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

[7] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. ACM SIGKDD Int'l Conf., ACM Press, 2004, pp. 168–177.

[8] J. Wiebe et al., "Learning Subjective Language," Computational Linguistics, vol. 30, no. 3, 2004, pp. 277–308.

[9] A. Abbasi and H. Chen, "Applying Authorship Analysis to Extremist-Group Web Forum Messages," IEEE Intelligent Systems, vol. 20, no. 5, 2005, pp. 67–75.

[10] A. Esuli and F. Sebastiani, "Senti-WordNet: A Publicly Available Lexical Resource for Opinion Mining," Proc.5th Conf. Language Resources and Evaluation (LREC-06), Evaluation and Language Resource Agency, 2006, pp. 417–422.

[11] B. Liu, Web Data Mining, Springer, 2007.

[12] A. Devitt and K. Ahmad, "Sentiment Polarity Identification in Financial News: A Cohesion-based Approach,"Proc. 45th Ann. Meeting Assoc. Computational Linguistics, ACL Press, 2007, pp. 984–991.

[13] K. Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis," Proc. IEEE 24th Int'l Conf. Data Eng. Workshop (ICDEW 2008), IEEE Press, 2008, pp. 507–512.

[14] A. Fahrni and M. Klenner, "Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives," Proc.AISB Symp. Affective Language in Human and Machine, Soc. Study of Artificial Intelligence and Simulation of Behavior Press, 2008, pp. 60–63.

[15] S. Li and C. Zong, "Multi-domain Sentiment Classification," Proc. Assoc.Computational Linguistics, ACL Press, 2008, pp. 257–260.

[16] Yee W. LO1, DEBII and Vidyasagar POTDAR2, DEBII "A Review of Opinion Mining and Sentiment Classification Framework in Social Networks", 2009, 3rd IEEE International Conference on Digital Ecosystems and Technologies.

[17] Yoonjae Jeong, Youngho Kim, Seongchan Kim, and Sung-Hyon Myaeng, Hyo-Jung Oh ,"Generating and Mixing Feature Sets from Language Models for Sentiment Classification", IEEE NLP-KE,2009

[18] Zhu Zhang, University of Arizona, Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications, 2008,IEEE, NLP and the Web.

[19] Dongmei Zhang1,2, Shengen Li1, Cuiling Zhu2 and Xiaofei Niu1,2, Ling Song1,"A Comparison Study of Multi-class Sentiment Classification for Chinese Reviews", 2010, Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010).