EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 11, November 2015)

An Improvised Approach for Utilizing Sentiment Analysis for Topic Detection

Sheetal Lohare[1], Prof. Kavita Bhosale[2]

[1]M.Tech-Student, [2]Assistant Professor, Department of Computer Science and Technology, Maharashtra Institute of Technology, Aurangabad, India

***Abstract*- Due to the vast growth and emergence of the consumer generated media (CGM) on internet such as websites, blogs, forums and news articles, ecosystem of corporations has changed significantly. Customers, retailers are tremendously vocal about reviews and insight of companies, their brands, products and services offered on the web. The reviews of customers are really important to gain huge number of customers. So recently, sentiment analysis of online customer reviews has emerges as a very interesting research topic. Proposed sentiment analysis put emphasize to categorize web comments into positive, neutral, and negative categories and identifies the topics which are closely interrelated with the positive and negative reviews available on web. These approaches when combined with sentiment classification helps in decision making.**

***Keywords*- Sentiment analysis, opinion mining, topic detection, Pointwise Mutual Information (PMI) value, word support**

I. INTRODUCTION

Rapid growth and availability of consumer generated media (CGM) and user-generated content such as websites, blogs, forums, news article and message boards place not only immense opportunities but also some hazards on today's activities. Contents on available on forums, blogs and news articles can be used to evaluate customer's opinion, view about products and services delivered by vendors. This online word introduces a new and significant source of information which is helpful in business intelligence and marketing [2]. Reviews of consumers are very useful to other possible consumers, manufactures and retailers to examine the views of consumers. Reviews about specific product and services are also helpful in decision making. A review generates new innovative opportunities and competitive advantages [16]. Once the number of opinions of consumers increases, it's very difficult to analysis the opinions of preceding customers about various aspects of products and services with a manual analysis.

If CGM and user-generated content are ignored, then companies could fall in considerable risks if some problems are not properly handled early and efficiently.

As the CGM information spread rapidly on internet, one cannot render the bad publicity of specific product or service [3]. Efficient analysis and summarization approaches are required to visualize earlier positive and negative opinions, reviews about specific characteristics or aspects of products and services. Therefore it is extremely advantageous to construct new analytical model that will leverage CGM content to understand consumer opinions.

As sentiment analysis term analyzes the opinions of consumers it is also called as opinion mining [15]. In past few years, opinion mining has gain much more attraction for evaluate online customer reviews using data mining techniques and natural language processing [6]. Sentiment analysis is a type of text analysis [1] which mostly includes text mining and computational intelligence.

In Sentiment analysis, we classify comments of consumers in three categories as positive, negative, and neutral categories. Analysis of comments using these categories is helpful, but it does not provide the detail insight about the actual reason behind the comments. Proposed method tackles with this issue by combining a unique sentiment classification approach with a topic detection approach. Proposed method helps to find out terms which are extremely interrelated with distinct sentiment classification types. The overall solution determines the sentiment about a given topic as well as exposes the possible origin reasons of the sentiments.

II. RELATED WORK

Existing techniques used for analysis of consumer's opinions concentrates on sentiment classification, whose goal is to distinguish opinions, views of users about any specific product and categorize them into positive, negative and neutral categories. There are two major categories used for sentiment classification are as follows:

A. Semantic-based approach

Sentiment based classification approach is depends on group of opinions words which further forms the structure like sentiment dictionary or a large-scale knowledge base [5] to allocate sentiments to individual documents.

Opinions words used for sentiment based classification approach implies sentiment words which carries positive or negative sentiment such as Excellent, Good, Well, Worst, etc.

As the sentiment word collection is recognized, sentiment classification is carried out by examining average semantic orientation of all sentiment words present in each document. Only the thing is that, proper techniques should be adopted to generate appropriate semantic word base. Manual construction [4], semiautomatic construction [10], [11] and automatic construction [7] [8] are the techniques used for the construction of semantic word base. These semantic approaches are often adaptive and easy to use. Generating the baseline sentiment word base can be challenging task.

B. Learning-based approach.

The learning- based approaches leverage the manually labeled documents as the training set, and then for performing sentiment classification carries out learning methods as Naïve Bayes, Maximum Entropy and SVMs [12], [14]. Comments can be represents by Words, syntactic relations and n-grams. But learning-based approaches possess some disadvantages as follows:

1. Construction of labeled documents is challenging task.
2. This approach may not be more adaptive to work across distinct data sets or domains.

Following are some definitions used in the sentiment analysis process-

Snippet:

A snippet defined as a small text segment which lies around the specific keyword described in a given text document. Sentence boundaries or the number of words are used to describe the text segment. Generally, snippets are assembled in the region of core keywords like corporation or brand names of products. The process of snippetization is required for scrutinizing web content as contents on the web may often noisy in nature. Some web content contains various different topics in one document even if only some of them may be relevant with the analysis subject. Snippets permit users to concentrate on the relevant text segments. This is especially important to sentiment analysis, since sentiment analysis of the overall document is likely to bias the opinion of the concerned subject, which on-topic snippet-based sentiment analysis could be much more meaningful. Usually 1 or 2 pre and post sentences are sufficient to construct each snippet.

Semantic Dictionary:

Sentiment analysis consists of a semantic-based classification technique.

For sentiment classification, proposed method implies two categories of semantic dictionary as domain-specific and domain-independent.

Words which possesses general sentiment meanings are placed in domain-independent dictionaries, e.g., “good” and “bad”.

Words with different meanings in different domains are placed in domain-specific semantic dictionaries.

Domain Dictionary:

The term Domain dictionary is extensive form of sentiment dictionary. Domain dictionary consists of common words specific to a given domain irrespective of whether words are sentiment words or non-sentiment words.

Sentiment Topics:

Sentiment topics describe the overall background or associated information lies behind each concerned sentiment in document. In proposed system, each topic will be described through a set of representative words.

III. PROPOSED SYSTEM

The architecture of proposed sentiment analysis is based on the two key components as the sentiment classification component and the sentiment topic recognition component.

Role of the sentiment classification component is to calculate the overall polarity of each snippet and generate sentiment taxonomy according to the polarity of snippets. Depending on the results generated by sentiment classification component, the topic detection component recognizes the most appropriate information related to each sentiment category.

The entire process is carried out as follows:

1. Assemble the entire relevant web comments on specific objects from content repository like blogs, message boards, and news articles etc, and construct the web content data warehouse.
2. To scrutinize given set of subjects as names of products or brands, extort all snippets from the data warehouse related to specific subject.
3. Count the sentiment score for each snippet.
4. Categorize snippets into different sentiment types depending on their sentiment scores and generate the sentiment taxonomy.

5. Determine the most relevant topics to given sentiment category.

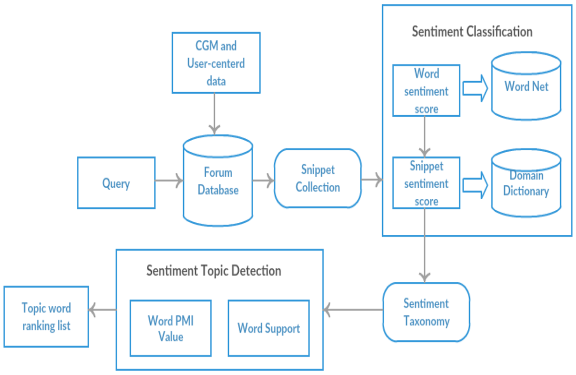

Figure 1: Proposed sentiment analysis framework

A. Key sentiment analysis components

1. Sentiment classification component.

Proposed classification technique is depends on semantic-based methods. These methods calculate the positive or negative polarity of word as well as compute the degree of sentiment for each expressed words and provides sentiment scores to each words depending on definitions.

These techniques are the foundation of the classification approaches. Generally, proposed sentiment classification method contains four sub-steps as follows:

1) Construct sentiment lexicon;
2) Compute sentiment of individual word;
3) merge all words in the snippet to generate the final sentiment score for the snippet;
4) Generate sentiment classes depends on the snippet scores.

2. Sentiment topic words recognition component.

Sentiment classification summarizes views of peoples about specific product but does not reveal the actual reasons behind the opinion. Goal of sentiment topic recognition component is to attempt such problems. In proposed sentiment topic recognition component, the significance of each topic word is calculated by two methods as Pointwise Mutual Information (PMI) value and word support.

PMI technique is used to assess the association used in information theory and statistics. PMI value between two distinct random variables determines their dependence. If the value between two variables is zero then that variables are independent, if value is higher than zero then two variables are completely associated and if value is much smaller than zero then variables are complementary to each other [16]. PMI can differentiate the association between variables, but it is always biased towards irregular words. Word support is used to balance the evaluation of association.

B. Sentiment topic detection algorithm

1. Sentiment based taxonomies

Sentiment analysis initiated by construction of efficient sentiment taxonomy. We make use of a statistically based technique to evaluate sentiment analysis which does not assume or attempt to determine sentiment for any particular subject or object. We calculate the positive or negative relative score of the sentiment expressed by the words in each snippet. The resulted relative score is further used to categorize snippet into positive, negative, neutral categories which required for generating sentiment taxonomy.

Establish positive/negative words list:

For generating sentiment lexicon, first we have to construct list of positive and negative word. To construct the positive/negative words lists we use two natural language processing (NLP) resources as The Inquirer database 2 and WordNet. The Inquirer database consists of more than 4,000 unique words and for each word database defines approximately 200 Boolean attributes. Attributes of words play an important role to decide whether the word is used in positive sense or negative. Generally, words in Inquirer database are adjectives. WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Each nouns, verbs and adjectives are categorized into synonym sets. According to Inquirer, if the majority of synonyms of a word are fall under positive category then we consider the original word as positive. Similarly, if the majority of synonyms of a word are fall under negative category then we consider the original word as negative.

Establish degree of sentiment:

To calculate relative sentiment score between two distinct posts which contains both positive and negative words to express sentiment, we compute the degree of each positive and negative word.

This is done with the help of WordNet dictionary and by calculating the occurrence of positive minus negative words in the definition. To carry out the normalization process, divide the sum by the total number of definitions. The occurrence of the word itself in its own definition is counted only once, where as other positive or negative words can be counted multiple times. To enhance analysis, only adjectives definitions are used and other part of sentence is ignored. This computation provides the relative amount of sentiment of each individual word.

Expand to incorporate all words in domain that are defined in WordNet:

The technique used to compute score of negative/positive words lies in original wordlists may be used to score any word consists in Word-Net dictionary. As dictionary contains positive and negative words in its definition, each words n dictionary may possesses both a positive and a negative impact. This technique will hold less impact of sentiment score on the words in the dictionary than the words in a positive and a negative impact. Only words in WordNet are considered in sentiment classification. Words other than WordNet are strictly prohibited for sentiment classification.

Score snippets and partition into quintiles:

The snippet sentiment score is computed by the summation of the positive score of sentiment word for a given snippet, minus the summation of the negative sentiment word scores divided by the square root of the entire length of the snippet.

10.0 F *(*P* - *N*) / Math.sqrt (*snippet*.length ())

Where P and N stands for the accumulation positive and negative score of all words in snippet respectively. The method used for scoring is checked against human rated data and result obtained are seems to hold a high degree of correlation with structure of sentiment.

With this technique of scoring, we categorize data into five classes by applying sorting of snippets according to the sentiment score. There are two intense quintiles as positive and a negative class and rest of three middle quintiles are combined to form a single “neutral” class. So, Positive, Negative, and Neutral are three classes of the final sentiment taxonomy.

2. Sentiment topic words recognition

Two components are used to search sentiment topic words as word PMI value and word support.

PMI value is used to investigate the uniqueness of word against each sentiment category. The equation given below is used to evaluate PMI value of word w against the each sentiment category s.

$$\mathrm{PMI}\ (w,s) = \log(\ p(w, s) / (p(s) * (p(w) + 0.05))\ )$$

p(w, s) represents the co-occurrence between word (w) and category (s), p(s) describes the distribution of category s and p(w) calculates the distribution of word w in the entire snippet collection.

Factor p(s) can be ignored as it does not possesses any influence on words ranking for each category.

Word support evaluates the importance of word in each sentiment category. It is calculated as the following:

$$\mathrm{Freq}\ (w, s)= N(w, s) / \sum N(w, s)$$

Where, N(w, s) describes the number of word w in category s. PMI and word support techniques are used to calculate the significance of a word in every sentiment category from different aspects. By combining together they can be used to identify the related topic words effectively. If any one out of PMI or word support is ignored, it will negatively impact on the process of topic detection.

Following procedure is adopted for detection of related sentiment topical words:

1. Categorize text documents into positive, negative and neutral categories using sentiment classification techniques.
2. Classify all words in documents and filter them in accordance with stop words and sentimental words. Keep only non sentiment words as sentiment topical word candidates.
3. Compute frequency of all sentiment topical words across both single sentiment category and all categories to generate word supports.
4. Compute PMI value of all sentiment topical words.
5. Merge the frequency of the words in each category with its PMI value and choose the most frequent words with maximum PMI value as the final sentiment topic words.

IV. CONCLUSION

Proposed sentiment analysis combines sentiment classification techniques with sentiment topic detection scheme which can be helpful to the business analyst to understand the scope and reason behind sentiment.

In sentiment classification technique, relative sentiment expressed in the words in every snippet is calculated on a positive or negative scale. In accordance with the relative sentiment score snippet is categorize into positive, negative, neutral category. In sentiment topic detection technique, closely related topics behind each sentiment category are identified by using PMI and word support metrics. The combination of these two approaches identifies sentiments as well as exposes the implicit original reasons of the sentiment.

REFERENCES

[1] Pero Subasic and Alison Huettner, "Affect Analysis of Text Using Fuzzy Semantic Typing", Proceedings of the Tenth IEEE International Conference on Fuzzy Systems, 2001, pp. 483-496.

[2] S. Das and M. Chen. "Yahoo! for amazon: Extracting market sentiment from stock message boards", Proceedings of the 8th Asia Pacific Finance Association (APFA), 2001.

[3] J. Kamps and M. Marx, "Words with attitude", Proceedings of the First International Conference on Global WordNet, 2002, pp. 332-341.

[4] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 417-424.

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

[6] Tetsuya Nasukawa and Jeonghee Yi, "Sentiment AnalysisCapturing Favorability Using Natural Language Processing" Proceedings of the International Conference on KnowledgeCapture, 2003, pp. 70-77.

[7] Hugo Liu, Henry Lieberman and Ted Selker, "A Model of Textual Affect Sensing using Real-World Knowledge", Proceedings of the Seventh Conference on Intelligent User Interfaces, 2003, pp. 125-132.

[8] Peter D. Turney and Michael L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association", *ACM Trans. On Information Systems*, 2003, 21(4), pp. 315-346.

[9] Kushal Dave, Steve Lawrence and David M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of the Twelfth International World Wide Web Conference, 2003, pp. 519-528.

[10] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques", Proceedings of the IEEE International Conference on Data Mining (ICDM), 2003, pp. 427-434.

[11] M. Hu and B. Liu, "Mining and summarizing customer reviews", Proceedings of the 10th international conference on Knowledge discovery and data mining (KDD), 2004, pp. 168-177.

[12] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization", Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM), 2006, pp. 43–50.

[13] V. Ng, S. Dasgupta, and S. M. N. Arifin, "Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews", Proceedings of the 21st International Conference on Computational Linguistics (COLIN) and 44th Annual Meeting of the Association for Computational Linguistics (ACL), 2006. pp. 611-618.

[14] K. T. Durant and M. D. Smith, "Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection", *Lecture Notes in Computer Science*, Springer, vol. 4811, 2007, pp. 187-206.

[15] Keke Cai, Scott Spangler, Ying Chen!, Li Zhang, "Leveraging Sentiment Analysis for Topic Detection", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp 266-271, 2008.

[16] Ayoub Bagheri, Mohamad Saraee, Franciska de Jong, "Care more about customers: Unsupervised domain-independent aspectdetection for sentiment analysis of customer reviews", 2013.