ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 11, May 2016

International Conference on Innovations in Engineering and Technology (ICIET'16)

6th & 7th April 2016

Organized by

K.L.N. College of Engineering and Technology, Madurai, Tamil Nadu, India

Aspect - Based Sentiment Analysis Using Context - Aware Learning for Online Reviews

C.Mahalakshmi, A.M.Abirami

Department of Information and Technology, Thiagarajar college of Engineering, Madurai, India

Department of Information and Technology, Thiagarajar college of Engineering, Madurai, India

Abstract: To process the billions of daily social conversations across millions of sources require sophisticated streaming big data processing to churn through the huge volume of content. In order to filter the noise and spam and classify the information and intelligence, in turn, requires contextual semantic sentiment modeling to capture the intelligence through the complexity of online social discussions. Sentiment analysis is one of the techniques to capture the intelligence from Social Networks based on the user generated content. There are more and more researches doing about sentiment classification. Comparing researches about the document and sentence-level sentiment classification, aspect-based sentiment analysis researches are less. The aspect-based sentiment analysis is a fine - grained approach. The ultimate goal is to be able to generate summaries listing all the aspects and their overall polarity. The novel idea of this paper is to generate Context-Aware lexicon dictionary to capture contextual semantics polarity and highly relevant aspects are captured using Latent Semantic Analysis, which will improve the accuracy of aspect- based sentiment analysis

KEYWORDS: Sentiment analysis, aspect-based, Social Networks, latent semantic analysis, context semantic, context dependent

I. INTRODUCTION

The Internet becomes a part of everyone life because it connects people around the world quickly and easily, so the usage of internet increase rapidly. The Internet connects people through online activities like online shopping, online transaction, chatting, social media communication, micro blogging and blogging. The Social media is a forum which creates a great impact among the peoples because here people can share their thoughts and opinions without any hesitations from their standpoint. This provides businesses with the huge potential to engage and interact with the public easily. From a commercial point of view, people who use this social media can follow brands, business, access offers and promotions and also people can analyze products and services they want to buy. This is done by analyzing and evaluates different people opinions sharing in the social media. Blogs and Reviews plays a vital role in people product buying decision because More than half of social media people reading reviews and comments of a product before they make a purchase . Academia, research communities, and industries have been working on sentiment analysis to extract the opinion of people from this social media and analyze and evaluated the user views. This paper presents a survey on sentiment analysis on social media by using various approaches, techniques, algorithms to analyze the opinion of people in social media. Sentiment analysis (also called as Opinion Mining) is the task of identifying, extracting and classifying sentiment, opinions, and attitudes about various topics as uttered in textual input [2]. Sentiment analysis helps to achieve the goals like predict the market business intelligence, customer satisfaction, public awareness, and opinion about products [1].

In social networks, e-commerce is an important source to express and analyze the opinions, emotions and sentiments. In the current scenario shows that the producers and service providers can improve their quality and standard of products based on the reviews posted by customers. From a customer perspective, they make a decision like to buy this product or to subscribe this service from the opinions given by other customers [1].

Social Networks are used to share and express their experience, opinions in a variety of ways like blogs, forums, and product reviews. Social Networks sites contain huge and highly unstructured data like combining text, smiley, images, videos and animations. These makes the extracting opinion from the User Generated Content is more difficult.
The sentiment analysis is classified into document level, sentence level, and entity level. Document level analysis predicts the sentiment of a document [1].
Various steps are involved in extracting opinion from user generated content because the text coming from different sources in various format. Data acquisition and preprocessing are the most important task in Sentiment Analysis. Different Levels of Analysis are
Document level: The main task of this level is to classify whether positive or negative sentiment expressed in the whole document. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. For example, the system identifies whether the review states an overall positive or negative opinion about the product from the given product review. This level of analysis applicable for single entity not for multiple entities
Sentence level: the task of this level is to identify whether each sentences expressed sentiment orientation as positive, negative or neutral opinion. This is related to subjectivity analysis.
Entity and aspect level: Both the document level analysis and the sentence level analysis analyses only the opinion expressed in the document or sentence is positive, negative or neutral not an opinion holder. They do not find out what exactly people liked or not liked. Aspect level analysis is fine grained analysis. It is also known as feature level analysis.
There are mainly three approaches in sentiment analysis
1. Lexicon based - considers lexicon dictionary for identifying polarity of the text
2. Machine learning based approach - Needs to develop classification model, which is trained using pre-labeled dataset of positive, negative, neutral content.
3. Combined approach - Which uses lexicon dictionary along with pre-labelled data set for developing classification model.

The main task of Corpus-based approach is identifying the opinion words with context specific orientations. Its methods depend on syntactic patterns or co-occurrence with a seed list of opinion words to find other opinion words in a large corpus.

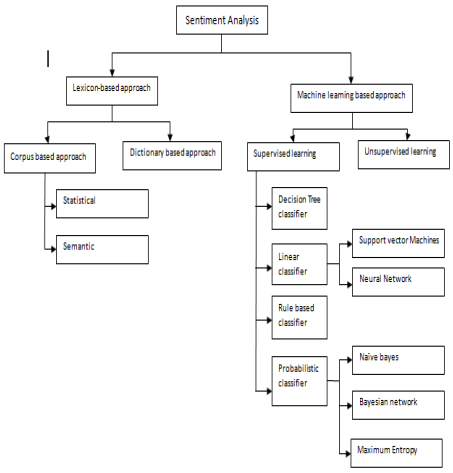

Fig.1 Sentiment Analysis Taxonomy Model

For example, consider the word "unpredictable", when it comes in movie review "The film is unpredictable", it becomes positive opinion. When it comes in software domain "Unpredictable program", it becomes negative opinion. In the first sentence, unpredictable is opinion word, the story is an implicit aspect. In the second sentence, unpredictable is opinion word, the program is an explicit aspect. Here the same word gives different polarity based on the context. In this paper proposes a more fine-grained aspect based approach using context-aware learning. Context-aware learning deals the problem of ambiguity by attempting to examine the opinion of the term in a given context.

II.LITERATURE SURVEY

Turney Turney [2] used a set of patterns of tags for collecting two-word phrases from reviews. They engaged PMI-Information Retrieval (PMI-IR) to examine semantic orientation of review by giving queries to a search engine, where "excellent" and "poor" have been considered as state line for positive and negative orientation words. They have taken 410 reviews on the different domains from Epinion.com for their experiment. The experiment result shows that they achieved the highest accuracy 84.0% for the smallest dataset and they achieved lowest accuracy 65.83 for movie

reviews dataset. In future work, the improvement can be done by tagging of sentences on the origin of whole or partial factor of discussion

Mullen and Collier [4] discussed a hybrid model comprising PMI-IR approach, Osgoodian semantic differentiation with WordNet, and syntactic-relation features and topic proximity. Osgoodian semantic gave the effectiveness (strong and weak), activity (active and passive), and the estimation factor (good or bad) for all adjectives. They tested with the dataset used in [3] and 100 reviews on media. The hybrid SVM with PMI/Osgood and lemmas feature selection produced the highest accuracy of 87% using 10-FCV. In future, to improve the performance classification, they Incorporate the domain context to AltaVista search results and feature selection based on dependency relation.

Chen et al. [5] discussed to classify sentiment of blogs on products. They have been performed Classification using a hybrid of semantic orientation with the back-propagation neural network. They discussed 4 different types of semantic orientation (SO) indexes as the input neurons such as SO-PMI (AND), SO-PMI (NEAR), latent semantic analysis and semantic association. Experiments were done on Blogs with data size 353 reviews (positive: 157, negative: 186), MP3 with data size: 579 (positive: 235, negative: 344), EC contains data size: 386 (positive: 249 negative: 137), Movie-001, and Movie-002 (positive: 500,negative : 500) and produced 87.5%, 73.5%, 81.2%, 75%, and 67.8% recall respectively. Their approach should have experimented on different corpora like Twitter, Facebook , Plurk etc applying various machine learning techniques.

Li et al. [6] improved frequency-based extraction and PMI-IR in order to examine product aspects. In order to find the frequent aspects from the dataset, they employed Apriori algorithm and useless rules are removed using compactness and redundancy rules. To confirm the number of relevant aspects using location based ordering and similarity i.e. PMI-IR based filtering. They have classified the examined aspects into aspects and non-aspects category. They used RCut with a threshold value of 1 on PMI-IR score. They tested with 120 reviews for nine categories of products, the average value of F-score is 73.3% and it is being outperformed Popescu and Etzioni [3]. In future, they may perform infrequent aspects extraction.

Jurgen[7], proposed a novel unsupervised method to bring a context-aware sentiment lexicon by utilizing the semi-structure of user-generated product reviews. To detect individual opinion expressions, determine their semantic orientation and relate them to specific product aspects mentioned in the review. An adjective is considered to be a proper opinion indicator if it either directly precedes an aspect. The likelihood - ratio test is a parametric hypothesis test. It is used to compare two different hypotheses H0 and H1 with regard to the fit to observed data. Hypotheses are formulated to decide whether a tuple exhibits negative orientation. In the case that neither positive nor negative orientation can be assigned, a tuple is classified as being neutral. They compared SentiWordNet vs SentiWordNet plus their approach

Ayoub Bagheri [9] proposed an unsupervised model which addresses the core tasks necessary to detect explicit and implicit aspects from review sentences in a sentiment analysis system. Their model differs from existing techniques in that it requires no labeled training data or additional information, not even for the initial seed information. Therefore, the model can easily be transferred between domains or languages.

Zheng-Jun Zha [10], has described Product aspect ranking is beneficial to a wide range of real-world applications. They have done an investigation and its usefulness in two applications, i.e. document-level sentiment classification that aims to examine a review document as expressing a positive or negative overall opinion, and review summarization. We perform extensive experiments to evaluate the efficacy of aspect ranking in these two applications and achieve significant performance improvements. The task of analyzing the sentiments expressed on aspects is called aspect-level sentiment classification. They have proposed probabilistic aspect ranking algorithm to examine the important aspects of a product from consumer reviews.

Shi [11] proposed work is to improve the performance of aspect extraction from online reviews. Extracting product aspects from online consumer reviews is a difficult task due to some unique characteristics of online reviews. They have proposed a method that extracts aspects based on frequent nouns and noun phrases. It adapts PMI-IR to the problem of aspects extraction and denotes the limitations of previous methods to improve the performance and generality of aspect extraction. They have used Frequency based mining and pruning, Order based filtering and Similarity-based filtering – using PMI – IR to solve the problem. In future, they plan to extract infrequent aspects to improve the recall of aspect extraction.

Hassan Saif [12], has described capturing the contextual sentiment for reviews. They used SentiCircle, a lexicon-based approach for sentiment analysis in twitter. It is different from the typical lexicon-based approach, which offers a fixed and static prior sentiment polarity of words regardless of their context, SentiCircle takes into account that the co-occurrence patterns of words in different contexts in reviews to capture their semantics and update their pre-assigned strength and polarity in sentiment lexicons according to Sentiwordnet that is being used for the prior sentiment.

III. PROPOSED WORK

A. General Idea

In this paper, the proposed method takes as input a list of product aspects and opinion word associated with that aspect derived from the free text reviews. The second task of our proposed system is Aspect- opinion pair Generation. Third task is Context-aware Sentiment lexicon Generation. Fourth task is polarity detection. Fifth task is Sentiment analysis

and Aspect Summarization using LSA. This system takes online consumer reviews for processing. The online review format differs website to website such as free – text format, pros and cons about product aspects, overall rating of the product. The system takes input as pros and cons and free text format.

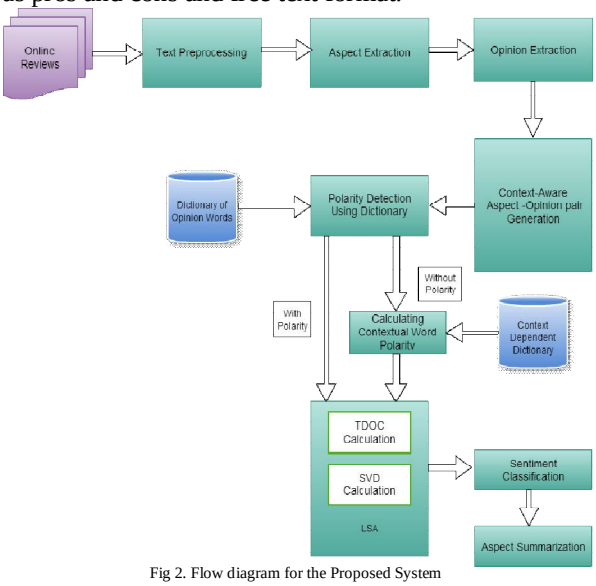

Fig 2. Flow diagram for the Proposed System

B. Preprocessing

After downloading the product reviews from the net, we separate pro and cons as a separate dataset and Free text reviews as separate file dataset. Preprocessing is done by using Stanford POS tagger. It tokenizes the review and returns the POS tags of the reviews along with the confidence score. After obtaining the tokenized and tagged reviews, we move to the next step of Preprocessing like filtering (extra characters like happpyyyy) ,questions (removing questions like how, what), removing special characters ([],(),{}) and stopword removal. The output of this step is to produce information - rich content.

C. Aspect Extraction

The system generates aspect vocabulary using pros and cons of consumer reviews which contains explicit aspects. Then we extract the aspects using frequent noun and noun phrases from the tagged reviews. After that the extracted frequent nouns and noun phrases are compared with aspect vocabulary. Then the system produces most frequently used aspect. Here pros denote the positive opinion and cons denotes negative opinion.

D. Aspect - Opinion Pair Generation

The subjective words also known as Opinion words (words that express the opinion). The earlier studies show that opinions are expressed by adjectives and verbs in the sentences. The opinions are obtained by the extraction of those adjectives and verbs. The Aspect opinion pairs are generated based on the opinion keywords. The nearest aspect of the opinion words are found and being analyzed to form the aspect- opinion pair.

E. Context - dependent lexicon generation

Context-aware learning deals the problem of ambiguity by attempting to examine the opinion of the term in a given context. After aspect-opinion pair generation, we find the polarity of the opinion word using the online dictionary which contains positive, negative and neutral score for every opinion word. If the polarity of the opinion words is neutral, then that words are identified as a context terms or ambiguous term. The system collects context terms and stores them in a context dependent sentiment lexicon. The number of co-occurring context terms in positive and negative documents acts as an indicator for the ambiguous terms' positive or negative charge.

For example "long battery life with high picture quality" and "This program takes a long time to execute". In this example, "long", comes positive in the first sentence and used in camera domain and negative in the second sentence and used in the software domain. So the same word (long) has a different meaning in different domains. "Amazing, great picture quality and low prize", and "camera have low resolution". In the first sentence, "low" is positive and it is used in camera domain. In the second sentence, "low" is negative and it also used in the same domain. These words are known as context-dependent words i.e. polarity of the word depends on the context. The context-dependent adjectives were classified as neutral by SentiWordNet and thus, so it did not consider as part of an opinion expression. The online consumer reviews contain context dependent opinion words. The proposed idea is to find the polarity of context-dependent words using the context dependent dictionary which contains context dependent lexicons and its score.

F. Polarity Calculation and Sentiment analysis Using LSA

Term degree of correlation

This feature represents the degree of correlation between a term A and its opinion term O. this will give how important the opinion term O is to Aspect A.

$$TDOC(A, O_i) = f(A, O_i) * log \frac{T}{T_{O_i}}$$

This step generates Correlation matrix. After that, a mathematical technique called singular value decomposition (SVD) is used to replace the original term degree correlation matrix with a much smaller matrix. As part of this process, unimportant words get discarded, and highly relevant or more important words are singled out. The new matrix can be used to place associated opinion and aspects into categories.

Prior sentiment score calculation

Each opinion term in the Aspect- Opinion pair (A, O_i) is assigned to a prior sentiment score based on its POS tags by using SentiWordNet Dictionary.

Negation

In the review, the opinion word O appears within the scope of a negation means its sentiment score is negated. For example, in the review ‘‘Picture quality is not good!’’, the term ‘‘good’’ is preceded by a negation. The original sentiment score is negated .From the above example, sentiment score of the word good is 0.75 in the SentiWordNet , we use this score negated (0.75) i.e., -0.75.The negation words are collected from the General Inquirer under the NOTLW category.

G. Aspect Summarization

If LSA(A, Oi) * PriorSentimentScore(Oi) comes positive, we’ll increase the positive count of the corresponding aspect. If it comes negative, we’ll increase the negative count of the corresponding aspect.

IV. EXPERIMENT

A. Data Collection

The dataset contains consumer reviews of 5 products. There are 3612 reviews in total and around 600 reviews for each product an average. The datasets are crawled from amazon.com and CNet.com . F1-measure was used as the evaluation metric to identify the aspect and classify the aspect sentiment. It is a combination of precision and recall, as F1 = 2 x precision x recall/(precision + recall) to evaluate the performance of product reviews.

B. Extracting aspect & Sentiment analysis

For the Positive and negative reviews, we identify the aspects by extracting the frequent noun terms in the reviews. From the previous study we know that aspects are noun or noun phrases. So we can obtain highly accurate aspects by extracting frequent noun terms from the reviews. Generally, an opinionated expression is associated with the aspect and the document contains at least one sentiment term in the sentiment lexicon, and it is the closest one to the aspect within the context distance of 5. The learned sentiment classifier is then leveraged to determine the opinion on the aspect. Sentiment score is calculated by TDOC and SVD combined LSA method.

C. Performance Measure

To evaluate the experiment result, this paper uses four indexes. They are Accuracy, Precision, Recall and F1 measure. The accuracy is used to estimate the whole effect of aspect extraction. F1 measure was used to evaluate the comprehensive effect of Precision and Recall. These performance measures can be calculated through table 1. Y means the review contains the aspect. X means the review does not contain the aspect.

TABLE 1 . Confusion Matrix

Results of the Experiments		
	Predicted X	**Predicted Y**
Actual X	A	B
Actual Y	C	D

$$\text{Accuracy} = \frac{a + d}{a + b + c + d}$$

$$\text{Precision} = \frac{d}{b + d}$$

$$\text{Recall} = \frac{d}{c + d}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

V. CONCLUSION AND FUTURE WORK

In this research, we study aspect based sentiment analysis for online consumer reviews. From this study shows that ambiguous terms also present in the reviews. The polarity of the ambiguous terms are depends on the context, i.e., the same word can act positive in domain and negative in domain. Normally those types of words are considered as neutral. In this paper, we proposed lexicon based context depend model for detecting context dependent terms and improve the accuracy of the Aspect based sentiment analysis for Product reviews using context aware learning and also we are

using LSA, which produce more relevant aspects . The proposed model able to handle two major problems: Context dependent words and Aspect summarization. In future we will include implicit aspect in our work.

REFERENCES

[1]. B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.
[2]. P. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In: Proceedings of the 40th annual meeting on association for computational linguistics ACL'02 (417–424), Stroudsburg, PA, USA: Association for Computational Linguistics, 2002.
[3]. A. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005, pp . 339–346.
[4]. T. Mullen, N. Collier, Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, 412-418), 2004, July.
[5]. L.-S. Chen, C.-H. Liu, H.-J. Chiu, A neural network based approach for sentiment classification in the blogosphere, Journal of Informetrics 5 (2011) 313–322.
[6]. S. Li, L. Zhou, Y. Li, Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures, Information Processing and Management 51 (2015) 58–67.
[7]. Jurgen Bro, Heiko Ehrig ,Generating a Context-Aware Sentiment Lexicon for Aspect-Based Product Review Mining, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010
[8]. Wenhao Zhang, Hua Xu , Wei Wan , Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis, Elsevier Expert Systems with Applications 39 (2012) 10283–10291
[9]. Ayoub Bagheri , Mohamad Saraee , Franciska de Jong Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews, Elsevier , Knowledge - Based Systems 2013
[10]. Zheng-Jun Zha, Jianxing Yu, Jinhui Tang,Meng Wang, Member, and Tat-Seng Chua ,Product Aspect Ranking and Its Applications, IEEE transactions on knowledge and data engineering, vol. 26, no. 5, may 2014 1211
[11]. Shi Li, Lina Zhou , Yijun Li , Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures, Elsevier, Information processing and Management 58-67 , 2015