IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS(ICECS '2015)

Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm

A.Jeyapriya (P.G Scholar)
Department of Computer science and Engineering
Kongu Engineering College
Erode, Tamilnadu, India.
dauntlessjeya@gmail.com

C.S.Kanimozhi Selvi (Associate Professor)
Department of Computer science and Engineering
Kongu Engineering College
Erode, Tamilnadu, India.
kanimozhi@kongu.ac.in

***Abstract*— Social media is emerging rapidly on the internet. This media knowledge helps people, company and organizations to analyze information for important decision making. Opinion mining is also called as sentiment analysis which involves in building a system to gather and examine opinions about the product made in reviews or tweets, comments, blog posts on the web. Sentiment is classified automatically for important applications such as opinion mining and summarization. To make valuable decisions in marketing analysis where implement sentiment classification efficiently. Reviews contain sentiment which is expressed in a different way in different domains and it is costly to annotate data for each new domain. The analysis of online customer reviews in which firms cannot discover what exactly people liked and did not like in document-level and sentence-level opinion mining. So, now opinion mining ongoing research is in phrase-level opinion mining. It performs finer-grained analysis and directly looks at the opinion in online reviews. The proposed system is based on phrase-level to examine customer reviews. Phrase-level opinion mining is also well-known as aspect based opinion mining. It is used to extract most important aspects of an item and to predict the orientation of each aspect from the item reviews. The projected system implements aspect extraction using frequent itemset mining in customer product reviews and mining opinions whether it is positive or negative opinion. It identifies sentiment orientation of each aspect by supervised learning algorithms in customer reviews.**

***Keywords*— aspect based opinion mining, frequent itemset mining, sentiment orientation .**

I. INTRODUCTION

Data mining research has successfully shaped numerous methods, tools, and algorithms for handling huge volume of data to solve real world problems. The key objectives of the data mining process are to effectively handle large scale data, mine actionable rules, patterns and gain insightful knowledge. The explosion of social media has created extraordinary opportunities for citizens to publicly voice their opinions. Because societal media is widely used for diverse purposes, huge content of user created data exist and can be made an accessible for data mining. Recent researches in data mining focus on opining mining.

A. Opinion Mining

Opinion mining is extracting people's opinion from the web. It analyzes people's opinions, appraisals, attitudes, and emotions toward organizations, entities, persons, issues, actions, topics, and their attributes in Liu [1]. Opinion is quintuple(e_j , a_{jk} , so_{ijkl}, h_i , t_l) where e_j is target entity , a_{jk} is aspect of entity, h_i is opinion holder, t_l is the time when the opinion is expressed , so_{ijkl} is sentiment orientation of opinion holder h_i on feature a_{jk} of entity e_j at time t_l . Users express their opinions about products or services they consume in blog posts, shopping sites, or review sites. It is useful for both the consumers as well as for the producers to know what general public think about a particular product or service. Sentiment analysis and opinion mining aim to automatically extract opinions expressed in the user-generated content. There are many social media sites reporting user opinions of products in many different formats. Monitoring these opinions related to a particular company or product on social media sites is a new challenging one. Opinion mining tools allow businesses to understand new product opinion, product sentiments, brand view and reputation management. These tools help users to perceive product opinions or sentiments on a global scale. Supervised learning algorithms that require labeled data have been successfully used to build sentiment classifiers for a given domain. However, sentiment is expressed differently in different domains, and it is expensive to interpret data for each novel domain.

B. Levels of opinion Mining

Opinion mining is a method of tracking feel of the civic about a particular item, company, events and issues. This organization analyzes which part has opinion expressing, who wrote the opinion and what is being commented online reviews. There are three general categorizations for opinion mining tasks: document-level, sentence-level, and phrase-level in Liu [1]. Document-level tasks are mainly formulated as classification problems where the input document should be classified into a few predefined categories. In subjectivity classification, a document is classified as subjective or objective. In sentiment classification, a subjective document is classified as positive, negative, or neutral. Opinion helpfulness prediction classifies an opinion as being helpful or not helpful. Finally, opinion spam detection classifies opinions as spam and

not spam. Sentence-level opinion mining is performed at the sentence level. In opinion search & retrieval and in opinion question answering, sentences are usually retrieved and ranked based on some criteria. Opinion summarization aims to select a set of sentences which summarizes the opinion more accurately. Finally, opinion mining in comparative sentences includes identifying comparative sentences and extracting information from them. Phrase-level opinion mining is also known as aspect based opinion mining. It performs finer-grained analysis and directly looks at the opinion. The goal of this level of analysis is to discover sentiments on aspects of items. Aspects that are explicitly mentioned as nouns or noun phrases in a sentence are called as explicit aspects. e.g., 'resolution' aspect in the review sentence "The resolution of this camera is nice". Implicit Aspects are not explicitly mentioned in a sentence but are implied, e.g., 'price' in the sentence "This camera is so expensive."

Mining opinions at the document-level or sentence-level is useful in many cases. However, these levels of information are not sufficient for the process of valuable decision-making (e.g. whether to buy the product). For example, a positive review on a particular item does not mean that the reviewer likes every aspect of the item. Likewise, a negative review does not mean that the reviewer dislikes everything. In a typical review, the reviewer usually writes both positive and negative aspects of the reviewed item, although his general opinion on the item may be positive or negative. In fact, document-level and sentence-level opinions cannot provide detailed information for decision making. To obtain such information, a finer level of granularity is needed. Hence, the proposed method focused on aspect based opining mining in which concentrates on explicit aspects. Section IV contains the proposed idea and techniques used. Section V shows the experimental results.

II.RELATED WORKS

The automatic analysis of user generated contents such as reviews, online news, blogs and tweets can be extremely valuable for tasks such as mass opinion estimation, corporate reputation measurement, political orientation categorization, stock market prediction, customer preference and public opinion study. Liu et al [2] proposed a method to summarize all the customer reviews of a product. It focused on mining product features on reviews by user commented content. The drawback is that there is no group features according to the strength of the opinions.

The projected system focused an approach called Dynamic Adaptive Support Apriori in Kanimozhi Selvi et al [3] to calculate the minimum support for mining class association rules and to build a simple and accurate classifier.

In sentiment classification, a classifier is trained using labeled data, annotated from the domain in which it is applied. Pang et al [4] examined whether it is sufficient to treat sentiment classification simply as a special case of topic-based categorization or whether special sentiment-categorization methods need to be developed. This approach used three standard algorithms: Naive Bayes classification, maximum entropy classification, and support vector machines (SVMs) for sentiment classification. In topic-based classification, all three classifiers have been reported to achieve accuracies of 90% and above for particular categories.

Turney [5] measured the co-occurrences between a word and a set of manually selected positive words (e.g., good, nice, excellent and so on) and negative words (e.g., bad, nasty, poor and so on) using pointwise mutual information to compute the sentiment of a word.

In Kanimozhi Selvi et al [6] proposed an approach to obtain the frequent itemsets involving rare items by setting the support thresholds automatically.

Kanayama et al [7] proposed an approach to build a domain-oriented sentiment lexicon to identify the words that express a particular sentiment in a given domain. By construction, a domain specific lexicon considers sentiment orientation of words in a particular domain. Therefore, this method cannot be readily applied to classify sentiment in a different domain.

Ding et al [8] focused on customer reviews of products. In particular, the author reviewed the problem of determining the semantic orientations (positive, negative or neutral) of opinions expressed on product features in reviews. So, the author proposed holistic approach that can accurately infer the semantic orientation of an opinion word based on the review context. It provided a new function which is used to combine multiple opinion words in the same sentence.

In Pang et al [9] focused on the methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact based analysis. This paper includes a material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion oriented information access services gives rise to. To facilitate future work, a discussion of benchmark datasets is also provided.

Ramage et al [10] introduced Labeled LDA, a topic model that constraints Latent Dirichlet Allocation by defining a one–to–one correspondence between LDA's latent topics and user tags. This allows Labeled LDA to directly learn word tag correspondences. Labeled LDA outperforms SVMs by more than 3 to 1 when extracting tag specific document snippets.

Zhang et al [11] focused on mining features. Double propagation works well for medium-size corpora. However, for large and small corpora, it can result in low precision and low recall. To deal with these two problems, two improvements based on part-whole and "no" patterns are introduced to increase the recall. It can rank feature candidates by feature importance which is determined by two factors: feature relevance and feature frequency.

Daume et al [12] proposed a semi-supervised (labeled data in source, and both labeled and unlabeled data in target) extension to a well-known supervised domain adaptation approach. This semi-supervised approach to domain adaptation is extremely simple to implement, and can be applied as a pre-processing step to any supervised learner.

In Edison et al [13] focused on aspect based opinion mining in the proposed system. Tourism product reviews are used as dataset in the system. Hotel and Restaurants corpus is taken as dataset to mine reviews in aspect level. The task of mining opinions and summarization is performed to provide customers a decomposed view of rated aspects.

III. PROBLEM DEFINITION

The people cannot analyze exact information in the document and sentence level opinion mining on customer reviews. Aspect level opinion mining is one of the solutions to problem. This gives fine detail information in aspect level. The goal of the task is to extract aspects on customer reviews. Mining opinions on online customer reviews whether it is positive or negative opinion. The projected system identifies the number of positive and negative opinions of each aspect in online reviews.

IV. PROPOSED SYSTEM

The architectural overview for our working model of the proposed system is shown in figure 4.1.

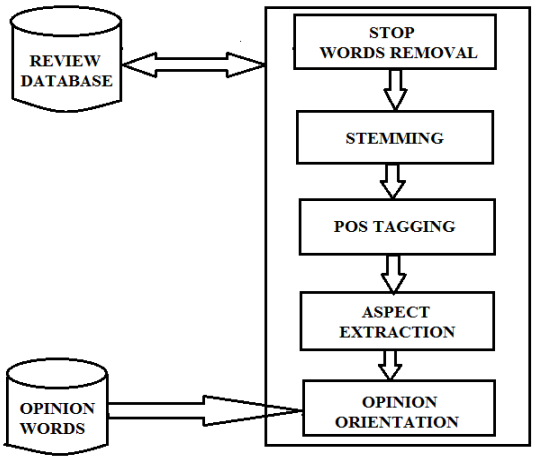

Figure 1. Working of Proposed System Architecture

The following section gives a detailed view of the proposed work. The proposed system uses customer reviews to extract aspect and mine whether given is positive or negative opinion. Each review is split into individual sentences. A review sentence is given as input to data preprocessing. Next, it extracts aspect in each review sentence. Stop word removal, stemming and pos tagging are data preprocessing. Sentiment orientation is used to identify whether it is positive or negative opinion sentence. Then it identifies the number of positive and negative opinions of each aspect.

```
Input: Online reviews
Output: aspects and sentiment orientation
Main procedure ()
        Data preprocessing ()
        Aspect Extraction ()
        Sentence and Aspect Orientation ()
End
Function Data preprocessing ()
        Stop words removal
        Stemming
        Pos tagging
End
Function Aspect Extraction (pos tagged input reviews)
            if word is in noun then
                extract (word)
            endif
            count numbers of each word
            set a minimum support count
            if aspect count < minimum support count
                display (word)
            else
                remove (word)
            endif
End
Function Sentence and Aspect Orientation ()
      Identify opinions using Naive Bayesian algorithm
End
```

Figure 2. Proposed Algorithm

C. Stop Word Removal

Most frequently used words in English are not useful in text mining. Such words are called stop words. Stop words are language specific functional words which carry no information. It may be of types such as pronouns, prepositions, conjunctions. Stop word removal is used to remove unwanted words in each review sentence. Words like is, are, was etc. Reviews are stored in text file which is given as input to stop word removal. Stop words are collected and stored in a text file. Stop word is removed by checking against stop words list.

D. Stemming

Stemming is used to form root word of a word. A stemming algorithm reduces the words "longing", "longed", and "longer" to the root word, "long". It consist many algorithms like n-gram analysis, Affix stemmers and Lemmatization algorithms. Porter stemmer algorithm is used to form root word for given input reviews and store it in text file.

E. POS Tagging

The Part-Of-Speech of a word is a linguistic category that is defined by its syntactic or morphological behavior. Common POS categories in English grammar are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. POS tagging is the task of labeling (or tagging) each word in a

sentence with its appropriate part of speech. POS tagging is an important phase of opinion mining, it is essential to determine the features and opinion words from the reviews. POS tagging can be done either manually or with the help of POS tagger tool. POS tagging of the reviews by human is time consuming. POS tagger is used to tag all the words of reviews. Stanford tagger is used to tag each word in an online review sentences. Every one sentence in customer reviews are tagged and stored in text file.

F. Aspect Extraction

Frequent itemset mining is used to find all frequent item sets using minimum support count. Here, every sentence is assigned as single transaction. Noun Words in each sentence is assigned as item sets for single transactions. Aspect extraction is implemented using figure 2. This algorithm first extracts noun and noun phrases in each review sentence and store it in a text file. Minimum support threshold is used to find all frequent aspects for a given review sentences. Aspects like pictures, battery, resolution, memory, lens etc. Then, the frequent aspects are extracted and stored in text file.

G. Sentence and Aspect Orientation

The proposed system first determines the number of positive and negative opinion sentence in reviews using opinion words. The positive and negative labels are collected labels in opinion words. Examples of positive opinion words are long, excellent and good and the negative opinion words are like poor, bad etc. And the next step is to identify the number of positive and negative opinions of each extracted aspect. Both sentence and aspect orientations are implemented using Naïve Bayesian algorithm using supervised term counting based approach. The probabilities of the positive and negative count are found according to the words using Naïve Bayesian classifier.

Naïve Bayesian algorithm

Steps are as follows:

1. The positive labels, negative labels and review sentences are stored in separate text file.
2. Split the sentence into the combination of words. It means first combination of two words and then single words.
3. First compare the combination of two words, if it matched then delete that combination from the opinion. Again start comparing of single word.
4. Initially, the probabilities of positive and negative count to zero [positive=0, negative=0]. The sentiment orientation algorithm is as follows:

```
if word is in opinion_words then
    mark(word)
    orientation ← Apply Opinion Word Rule
end if
if word is near a negation word then
    orientation ←Apply Negation Rules
end if
return orientation
```

Figure 3. Sentiment Orientation Algorithm

Opinion word rule in figure 3 gives that, if word is matched with positive opinion words then positive count get increment, or it is negative opinion word then negative count get increment.

In figure 3 *Negation rules* have a negation word or phrase which usually reverses the opinion expressed in a sentence. Two rules must be applied:

1. Negation Negative->Positive. This will increment positive count.
2. Negation Positive ->Negative. This will increment negative count.

After comparing all the words of the sentence, the found probabilities of the positive and negative counts are compared in the following manners.

a) If the probability of positive count is greater than the negative count, then the sentence or opinion is positive.

b) If the probability of negative count is greater than the positive count, then the sentence or opinion is negative.

c) If the probability of positive count minus probability of negative count is zero, then it is neutral.

Finally system identifies the number of positive and negative opinion of each extracted aspect in customer reviews.

V. EXPERIMENTAL SETUP

The following section describes the dataset used in our experiments and the results obtained.

H. Dataset Descriptions

The proposed system uses customer review dataset about a product effectively. A review is a subjective text containing a sequence of words describing opinions of reviewer regarding a specific item. Review text may contain complete sentences, short comments, or both. Product reviews are collected from websites like www.amazon.com, www.epinions.com and www.cnet.com. Each review in websites is assigned with a different rating like 0-5 stars, a review label and date, a reviewer name and location, a manufactured goods name, and the review content. Canon camera product reviews are used in the system. This dataset consists of product name and review text. Reviews are split into individual sentences. The details of the dataset used in the proposed system are shown in table 1 as follows,

Table 1. Corpus Details

Corpus	Canon Camera
Reviews	100
Total Sentences	400
Positive Sentence	231
Negative Sentence	108
Total Opinion Sentences	339
Opinion sentences(Percentage)	84.75%

I. Parameter For Evaluation

The performance of the system is evaluated. Precision, recall and F-measure are the parameters used in the system for evaluation. Precision is the measure of retrieved instances that

are relevant. Recall is the fraction of relevant instances that are retrieved. F-measure is a measure of test's accuracy. Precision, recall and F-measure are defined as follows,

$$\text{Precision} = \frac{|\text{ExtractedValues} \cap \text{TrueValues}|}{|\text{ExtractedValues}|} \quad (1)$$

$$\text{Recall} = \frac{|\text{ExtractedValues} \cap \text{TrueValues}|}{|\text{TrueValues}|} \quad (2)$$

$$\text{F} - \text{measure} = \frac{2 * \text{Recall} * \text{Precision}}{(\text{Recall} + \text{Precision})} \quad (3)$$

To calculate these measures, true values in reviews are identified manually. The proposed system mines the aspects and opinion (extracted values). Using this, precision, recall and F-measure are calculated for product customer reviews.

J. Results

Aspect extraction gives accuracy of 80.36% using frequent itemset mining. Sentiment orientation provides 92.37% of accuracy for given dataset. Precision, Recall and F-measure for aspect extraction and sentiment orientation are shown in figures 4&5 as follows,

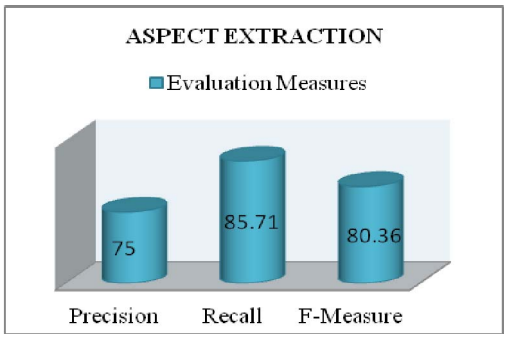

Figure 4. Parameters for Evaluation of Aspect Extraction

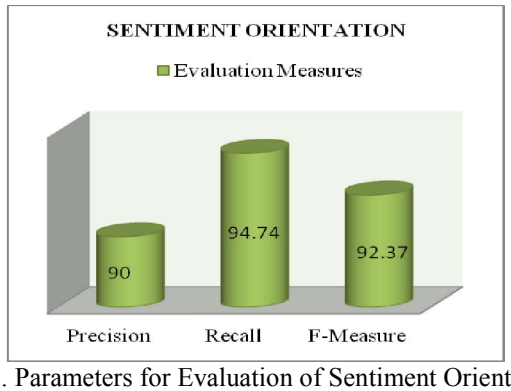

Figure 5. Parameters for Evaluation of Sentiment Orientation

VI. CONCLUSION AND FUTURE WORK

The proposed system extracts aspects in product customer reviews. The nouns and noun phrases are extracted from each review sentence. Minimum support threshold is used to find all frequent aspects for the given review sentences. Naïve Bayesian algorithm using supervised term counting based approach is used to identify whether sentence is positive or negative opinion and also identifies the number of positive and negative opinion of each extracted aspect. The number of positive and negative opinions in review sentences is estimated. Sentiment orientation gives a good accuracy. In future, it is proposed to summarize the aspects based on the relative importance of the extracted aspect. By using this, it is possible to analyze the customers interesting aspects on products.

ACKNOWLEDGMENT

Our sincere thanks to the experts who supported and guided us with their valuable domain knowledge.

REFERENCES

[1] Bing Liu (2012), 'Sentiment Analysis and Opinion Mining', Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

[2] Hu, Minqing and Bing Liu (2004), 'Mining opinion features in customer reviews', In Proceedings of the national conference on artificial intelligence, Vol.4, No.4, pp.755-760..

[3] Selvi, Kanimozhi, and A. Tamilarasi (2007), 'Association rule mining with dynamic adaptive support thresholds for associative classification', In Conference on Computational Intelligence and Multimedia Applications, International Conference, vol. 2, pp. 76-80.

[4] Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan (2002), 'Thumbs up?: sentiment classification using machine learning techniques', In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol.10, pp. 79-86.

[5] Turney and Peter D (2002), '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, pp. 417-424.

[6] Sadhasivam, Kanimozhi SC, and Tamilarasi Angamuthu (2011), Mining Rare Itemset with Automated Support Thresholds',Journal of Computer Science 7, vol 3, pp. 394-399.

[7] Kanayama, Hiroshi and Tetsuya Nasukawa (2006), 'Fully automatic lexicon expansion for domain-oriented sentiment analysis', In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.1-9.

[8] Ding, Xiaowen, Bing Liu and Philip S. Yu (2008), 'A holistic lexicon-based approach to opinion mining', In Proceedings of the 2008 International Conference on Web Search and Data Mining, Association for Computing Machinery, pp. 231-240.

[9] Pang, Bo and Lillian Lee (2008), 'Opinion Mining and Sentiment Analysis', Foundations and Trends in Information Retrieval, Vol. 2, No. 1/2,pp.1-135.

[10] Ramage, Daniel, David Hall, Ramesh Nallapati and Christopher D. Manning (2009), 'Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora', In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,Vol.1, pp.248-256.

[11] Zhang, Lei, Bing Liu, Suk Hwan Lim and Eamonn O'Brien-Strain (2010), 'Extracting and ranking product features in opinion documents', In Proceedings of the 23rd international conference on computational linguistics: Posters, Association for Computational Linguistics, pp. 1462-1470.

[12] Daumé III, Hal, Abhishek Kumar and Avishek Saha (2010), 'Frustratingly easy semi-supervised domain adaptation', In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, Association for Computational Linguistics, pp.53-59.

[13] Marrese-Taylor, Edison, Juan D. Velásquez and Felipe Bravo-Marquez (2014), 'A novel deterministic approach for aspect-based opinion mining in tourism products reviews', Expert Systems with Applications, Vol.41, No.17, pp.7764-7775.

[14] http://www.cs.uic.edu/~liub/