IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Aspect Level Opinion Mining on Customer Reviews using Support Vector Machine

Anju Joshi[1], Anubhooti Papola[2]

M. Tech Scholar, CSE Department, Uttarakhand Technical University, Dehradun, India [1]

Assistant Professor, CSE, Uttarakhand Technical University, Dehradun, India [2]

Abstract: With the rapid growth in ecommerce, reviews for popular products on the web have grown rapidly. Customers are often forced to wade through many online reviews in order to make an informed product choice. Scanning all of these reviews would be tedious, time consuming, boring and fruitless. It would be good if these reviews could be processed automatically and customers are provided with the limited generalized information. Opinion Mining plays a major role to summarize customer reviews and make it easy for online customers to determine whether to purchase the products or not.In this paper, we performed aspect level opinion mining on customer reviews using supervised learning algorithm. The proposed work performs the aspect-level opinion mining by extracting product aspects from reviews on ecommerce site and produces a summarized report of the most frequently discussed product aspects with regards to the number of appearances in positive, negative and neutral reviews. Here, frequent item set mining is used for aspect extraction and supervised learning algorithm (Support Vector Machine) is used to identify the number of the positive, negative and neutral opinion of each extracted aspect.

Keywords: Customer Reviews, Aspect Level Opinion Mining, Product Aspect Extraction, Supervised Learning Algorithm, Frequent Itemset Mining.

I. INTRODUCTION

Opinion mining is one of the most popular trends in today's world. It is the procedure by which information is extracted from the opinions, appraisal and emotions of people in regards to entities, events and their attributes [1]. It is the computational technique for extracting, classifying, understanding, and assessing the opinions expressed in various contents. Opinion mining is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing consumer brands, movie reviews, democratic electoral events and stock market. It has become one of the major parts in research, because of its enduring applications in marketing company, as they keep on exploring about their products, to initiate brand promotion; market segmentation and in framing new business strategies [2].

Nowadays E-commerce sites are gaining popularity across the world. People who buy the products online leave their comments or opinions about the products in E-commerce sites, which in turn help other customers to know about the product very well. So these online reviews not only help the customers to know about the products but also help the seller or manufacturer of the product to know what exactly the customers liked or disliked about the product. These online reviews plays very important role in choosing the product and to know about a particular product. But these reviews will not be in small amount, Reviews might be hundreds or even thousands. Customers cannot read so many reviews to come to conclusion whether to buy the product or not. Also reading few reviews will not be enough to finalize the product aspects. So it will be helpful for both customer and producer, if the product reviews are mined and presented in summarized manner [3].

In opinion mining a review can be determined in three separate levels: Document level, Sentence level, and Aspect level. In document level, whole review document is classified into either positive or negative class. Document level classification is considered as text classification problems. In subjectivity/objectivity analysis the review document is classified into predefined class (subjective, objective). When the subjective document is classified into positive or negative class then it is called opinion classification. In sentence level every sentence of a review is classified into positive or negative class [4]. Aspect level opinion mining gives summary and shows different aspects of the product. Aspect level opinion mining is fine-grained analysis and can provide help for both potential customers and manufacturers to know what aspect of the product the reviewers mostly like or dislike.

In this research paper, aspect level opinion mining is performed on product reviews. The paper implemented a system capable of extracting product aspects from reviews on ecommerce site and produces a summarized report of the most

DOI10.17148/IJARCCE.2017.6717

frequently discussed product aspects with regards to the opinion of the reviews they appear in. In our work**,** frequent itemset mining is used for aspect extraction and supervised machine learning algorithm (Support Vector Machine) is used to identify the number of the positive, negative and neutral opinion of each extracted aspect. The system is evaluated on the basis of three different evaluation measures. These measures include Precision, Recall, and F-measure The following section highlights the literature review. The next section shows the research methodology. Section IV shows results and analysis. And lastly the conclusion and future work section of the proposed work.

II. LITERATURE REVIEW

This section presents various work related to the opinion mining. There are basically two methods used for opinion classification, first method is supervised method and second method is unsupervised method. The supervised methods like naive Bayesian, SVM and maximum entropy classify the reviews based on machine learning. Second method is unsupervised method where the classification is based on certain syntactic patterns that are used to express opinions.

Bing Liu et al.[5] proposed two techniques i.e. a. novel framework for analysing and comparing consumer opinions of competing products and b. a new technique based on language pattern mining is proposed to extract product features from Pros and Cons in a particular type of reviews. Dataset are collected from epinions.com for an experiment. Author’s experiment results shows that their proposed method is very effective.

Minqing Hu et al. [6] proposed the method to mine and summarize all the customer reviews of a product based on data mining and natural language processing methods. Researcher performs this task in three steps (a.) Mining product features that have been commented on by customers, (b) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative and finally (c) is summarizing the results. Experimental results show that author’s purposed techniques very effective.

In [7] Qi Su et al. proposed a novel mutual reinforcement approach for feature-level opinion mining problem. This approach clusters product features and opinion words simultaneously and iteratively by fusing both their content information and sentiment link information. Under the same framework, based on the product feature categories and opinion word groups, they construct the sentiment association set between the two groups of data objects by identifying their strongest n sentiment links. Author’s purposed model provides a more accurate opinion evaluation. Experimental results make obvious researchers method outperforms the state-of-art algorithms.

Mining unstructured and ungrammatical reviews are discussed in [8]. In this paper, the author has summarized ungrammatical and unstructured user reviews based on Support Vector Machine (SVM). SVM is used for review classification and to conclude the summary of user’s opinion about the product.

Peter D. Turney [9] purposed an unsupervised learning algorithm for recognizing synonyms, based on statistical data acquired by querying a Web search engine. They evaluated using 80 synonym test questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym test questions from a collection of tests for students of English as a Second Language (ESL) used Point wise Mutual Information (PMI) and Information Retrieval (IR) to calculate the resemblance of pairs of words. For both tests, author’s algorithm obtains a score of 74%.

In [10] and [11] author focused on mining reviews using document-level. Here entire document is classified into positive, negative or neutral. Negation words were also handled in their work. Author had initially prepared seed list where some of opinion words are stored along with its polarity value. Opinion words that are extracted from the document are searched against the seed list. If not present then searching for synonyms of the word is done in WordNet. If found then word will be added to seed list and given same polarity. Here based on seed list and WordNet the polarity of opinion words is determined. The above mentioned sentiment analysis techniques discusses the classification and summarization of customer reviews.

Gamgarn Somprasertsri [12] dedicated their work to properly identify the semantic relationships between product features and opinions. They proposed an approach for mining product feature and opinion based on the consideration of syntactic information and semantic information by applying dependency relations and ontological knowledge with probabilistic based model.

Popescu et al. [13] developed an unsupervised information extraction system called OPINE, which extracted product features and opinions from reviews.

III. PROPOSED METHODOLOGY

In this proposed methodology, initially camera (canon G3) reviews are collected from ecommerce site (Amazon.com). The collected reviews are stored into text files in the form of documents and these documents are given as input for pre-processing stages. Pre-processing includes Stemming, Stop word Removal and POS tagging. Aspect extraction step will give frequent aspects from reviews. Opinion orientation is used to identify whether it is positive, negative or neutral opinion sentence based on aspects.

DOI10.17148/IJARCCE.2017.6717

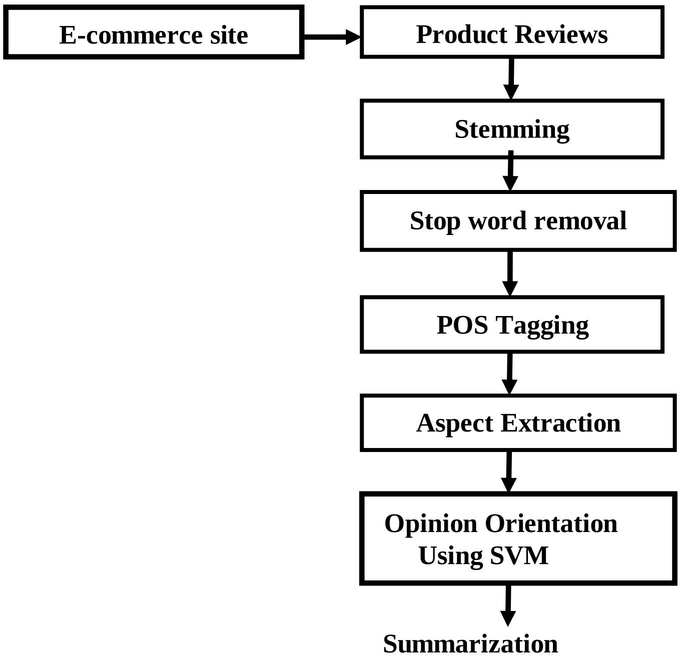


Figure 1 Architecture of proposed methodology

The details about the steps adopted in the methodology are described in the following subsections.
1. E-commerce Website: To determine the polarity of the reviews, based on aspects, large numbers of reviews are collected from the Web. There are lots of websites on the Internet such as Amazon, myntra, jabong, flipkart etc., where the large numbers of customer reviews are available.

2. Product Reviews: The proposed methodology 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. 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. In the thesis, reviews of Camera product **(Canon G3)** are collected to perform aspect level opinion mining. All the reviews are collected from Amazon. The dataset is stored in text file.

3. Data Pre-Processing: Before the implementation of aspect based opinion mining, data is pre-processed to get the understanding of data. There are various techniques have been used in data pre-processing like stemming, removal of stop words and parts of speech tagging.

- **Stemming:** In proposed system, stemming process is being used where a word will be reduced in its base or stem form for simpler processing. The stemmer algorithm is used for obtaining the root word from given review word and output of the stemming process is saved in the text file. A stemming algorithm reduces the words "longing", "longed", and "longer" to the root word, "long".
- **Stop word removal:** Stop words are words that have little meaning apart from other words and they are not needed in mining. The most popular stop words are “the,” “a,” “an,” “that,” and “those” and so on. These words rarely indicate anything about opinions. Reviews which are stored in text file are given to stop word removal algorithm. This algorithm removes stop words from the reviews by checking against stop word list. For example the sentence“The display is awesome” will be “display awesome” after stop word removal.
- **POS Tagging:** Part-of-speech tagging (or POS Tagging) is the task that assigns a part of speech tag to each word in a sentence like noun, adjective, adverb, verb etc. and identifies simple noun and verb groups. POS tagging of words is necessary to identify aspects and opinions expressed by the customer about the product in the review sentences. Aspects of a product are usually nouns and opinions are usually adjectives. Here Stanford POS tagger is made use, which tags all the online review sentences given to it. Every one sentence in customer reviews are tagged and stored in the text file.

For example, the sentence given to POS tagging after stop word removal is was “Battery life trash fast charge feature hardly ever works”. After POS tagging the sentence is Battery_NN life_NN trash_NN fast_JJ charge_NN feature_NN hardly_RB ever_RB works_VBZ.

4. Aspect Extraction: The aspect/feature of any particular product would be most of the times a noun or noun phrase. For that POS tagging has been used in the previous steps that detect words with tags like NNS (noun plural), NN

(Noun), NNP (proper noun singular) etc. But not all nouns are aspects; frequent item set mining is used to extract important aspects. Here, Apriori algorithm is used to find frequent aspect set. It is used to identify the most prominent aspects on which the customers have commented or expressed their opinions on, given a set of prospective aspects (denoted by the nouns) and the review dataset.

The algorithm counts the frequency with which words appear in different opinions, eliminating those less frequent. The end result is a subset of words called "frequent aspects" that have great chance of actually being a real aspect. Since a camera product is mined in the paper, some of the frequent aspects of camera are battery, picture, resolution, speed, price etc.

5. Sentence and Aspect Orientation: The proposed methodology first determines the number of positive, negative and neutral opinion sentence in reviews using opinion words .Words that encode a desirable state like excellent, good have a positive orientation, while words that represent undesirable states like poor, disappointing have a negative orientation. If sentence contains opinion words are taken as opinion sentence. 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, negative and neutral opinions of each extracted aspect. Both sentence and aspect orientations are implemented using Support vector machine (SVM).

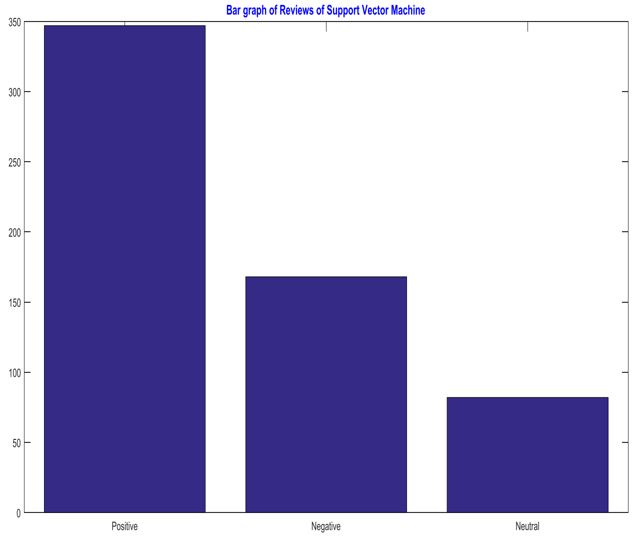
SVM is a machine learning classifier widely used for text categorization. The review text to be classified is converted into word vectors. SVM constructs a hyper plane using these vectors which separates data instances of one class from another. The unique property of SVM is that it can learn even if we provide huge data. SVM works well for text classification because it can handle large features. Another advantage of SVM is that it is robust when there is a small set of examples distributed over a large area.SVM has given reliable results in the past research in opinion mining. A review data classification is a two-step process. In the first footstep, a classification algorithm builds the classifier by "learning from" a training set made up of our corpus and their associated class labels. In a second step, the model is used for classification. A different set called test set is used to evaluate the correctness of the built model

IV. RESULT AND DISCUSSIONS

We conducted experiment by extracting Customer reviews of camera product (Canon G3) from the ecommerce website Amazon. The dataset consists of an approximately 597 sentences written by customers as opinionated reviews. Originally the dataset consist of 341 positive reviews, 179 negative reviews and 77 neutral reviews. The reviews are written as unstructured text files. The preprocessing is done on stored a reviews that is the stop words are removed from the review sentences. Then the reviews are split into sentences. The next step is tagging each and every word in the sentence to their respective parts of speech. This tagging is done using Stanford POS Tagger. POS tagging is necessary to identify Nouns.

Table 1: Number of positive, negative and neutral reviews using SVM

Positive	Negative	Neutral
345	162	90

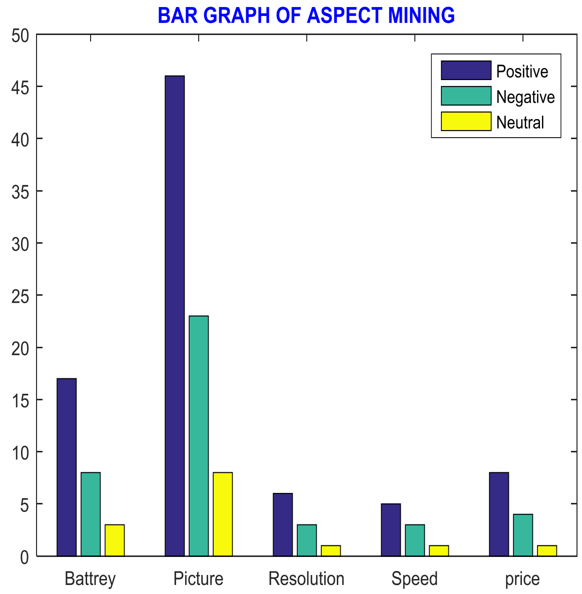


Graph 1: Number of positive, negative and neutral reviews using SVM

Nouns extracted are nothing but aspects and opinion words. Finally support vector machine (SVM) is used to classify the review text and to determine the number of the positive, negative and neutral opinion of each extracted aspect of the product (Canon G3 camera) taken into consideration.
SVM first determines the number of positive, negative and neutral opinion sentence in reviews and then identify the number of the positive, negative and neutral reviews based on each extracted aspect.

Table 2: Number of positive, negative and neutral reviews on camera aspects

Aspect Name	Positive	Negative	Neutral
Battery	17	8	3
Picture	46	23	8
Resolution	6	3	1
Speed	5	3	1
Price	8	4	1



Graph 2: Visual Aspect based summary of opinions on a Digital Camera

The aspect based summary obtained above can be easily visualized using a bar chart .The result of classification of reviews into positive, negative and neutral based on aspects for a product canon G3 is shown in graph 2 above. Here the aspect (picture) has maximum positive reviews when compare to other aspects and the aspect (speed) has minimum positive reviews when compare to other aspects. So this summarization based on aspect provides clear cut view of product to the customers and help them to decide whether to buy a product or not. And also helpful for manufacturer of the product to know which product aspect the customers liked or disliked, so in their future product they can improve their product quality.

The performance evaluation is calculated using Precision, Recall and F-measure. These measures are collected to determine effectiveness of the proposed work.

A) Precision

Precision is used to measure exactness. Precision is the number of reviews correctly labelled as positive divided on the total number that is classified as positive. In other words, it is the fraction of retrieved aspects that are relevant to search.

$$\text{Precision}= TP / TP+FP$$

B) Recall

Recall is a measure of completeness. Recall is the number of reviews correctly labelled as positive divided on the total number of reviews that truly are positive. In other words, it is the fraction of the relevant aspects that are successfully retrieved.

$$\text{Recall} = TP / TP+ FN$$

DOI10.17148/IJARCCE.2017.6717

C) F- measure

F-measure is the harmonic mean of precision and recall. This gives a score that is a balance between precision and recall. F-measure combines them into one score for easier usage. It is a combination of precision and recall.

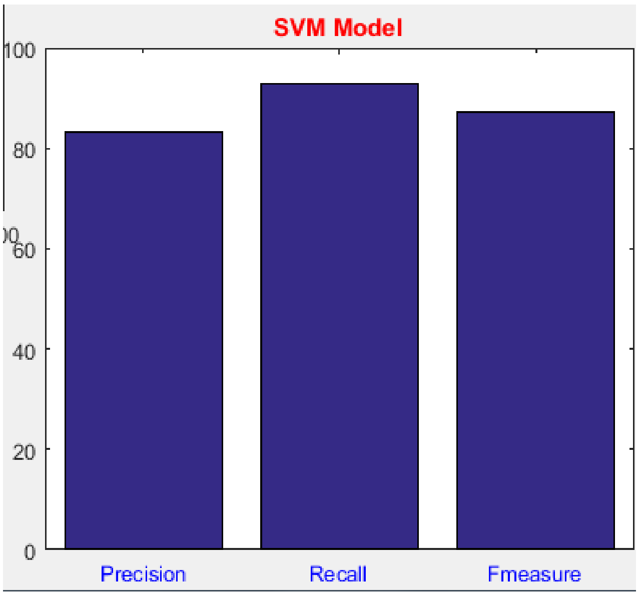
$$F\text{-measure}= 2*precision*recall / (precision+recall)$$

Where,

- True positives (TP) - number of reviews correctly labelled as belonging to particular class (positive/negative).
- False positives (FP) - number of reviews incorrectly labelled as belonging to particular class.
- False negatives (FN) - number of reviews were not labelled as belonging to the particular class but should have been labelled.

Table 3: Performance values using SVM

Method	Precision	Recall	F-measure
SVM	83.34	92.87	87.34



Graph 3: Performance values using SVM

Graph 3 plots the values of performance measures obtained by using support vector machine. From the graph it is evident that the precision is calculated as 83.34, recall is calculated as 92.87and F-measure is calculated as 87.34. The above graph shows an increase in all performance measures; consequently, the consistent results prove the validity of our proposed work.

V. CONCLUSION AND FUTURE SCOPE

In this paper, aspect level opinion mining is performed on customer reviews. An aspect-based summary of a large number of customer reviews of a product (canon camera G3) sold online is obtained. The proposed work extracts aspects using frequent itemset mining and the related customer opinions on the online (web) product domain based on each aspect is obtained using support vector machine. It proves to be an effective method that helps customers to analyse interesting aspects of the product. The proposed work allows summarizing the information obtained in order to provide a clear cut view of a product to the customers. Since the reviews collected were from the previous users, this summary helps the users to take a decision before buying a product online. Also this product summary helps the manufacturers to improve their product quality.

Many e-commerce sites contain product reviews, star rating and followers of reviews. As future enhancement, the proposed work can be extended to compare more than two products and product categories rather than just single product by considering all the factors which help in rating the product. Spelling correction component can also be added in the pre-processing of the reviews to improve the proposed System. Misspelled word identification will be helpful in extracting aspects correctly and relevantly. It will be helpful in improving the classification results.

DOI10.17148/IJARCCE.2017.6717

REFERENCES

[1] A.Jeyapriya, C.S.Kanimozhi Selvi "Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm" IEEE Sponsored 2nd International Conference On Electronics and Communication Systems (ICECS '2015).

[2] Venkata Rajeev P, Smrithi Rekha V "Recommending Products to Customers using Opinion Mining of Online Product Reviews and Features" International Conference on Circuit, Power and Computing Technologies [ICCPCT], 2015.

[3] Akshi Kumar, Teeja Mary Sebastian "Sentiment analysis: A perspective on its past, present and future" I.J. Intelligent Systems and Applications, 2015, 10, 1-14

[4] Anisha P Rodrigues, Dr. Niranjan N Chiplunkar "Mining Online Product Reviews and Extracting Product features using Unsupervised method " IEEE, 2016.

[5] B. Liu, M. Hu, and J. Cheng. Opinion observer: analysing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web, Japan, pages 342 - 351, 2005.

[6] Hu Minqing , Bing Liu, "Mining Opinion Features in Customer Reviews", In Proceeding of the National Conference on Artificial Intelligence, 2004, pp. 755-760

[7] Qi. Su M. Hu, "Mining and Summarizing Customer Reviews", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'04), USA, pp. 168-177, 2004.

[8] Taysir Hassan A. Soliman, Mostafa A Esmarsry ,"Utilizing Support Vector Machine in Mining Online Customer Reviews", ICCTA, 2012, pp. 192-196.

[9] P. D. Turney. Mining the web for synonyms, In Proceedings of the 12th European Conference on Machine Learning, EMCL '01, pages 491{502. Springer-Verlag, 2001.

[10] Richa Sharma, Shweta Nigam, Rekha Jain, "Opinion Mining of Movie Reviews at Document level", International Journal on Information Theory (IJIT), 2014, pp. 13-21.

[11] Vibha Soni, Meenakshi R Patel, "Unsupervised Opinion Mining from Text Reviews Using SentiWordNet", International Journal of Computer Trends and Technology(IJCIT), 2014, pp. 234-238

[12] Gamgarn Somprasertsri, Pattarachai Lalitrojwong (2008), A Maximum Entropy Model for Product Feature Extraction in Online Customer Reviews, King Mongkut's Institute of Technology Ladkrabang Bangkok 10520.

[13] Popescu, A. M., Etzioni: Extracting Product Features and Opinions from Reviews, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005, 339–346.

DOI10.17148/IJARCCE.2017.6717