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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Extracting and Mining Customer Reviews

Sapthami R [1], Anisha P Rodrigues [2]

PG Scholar, Department of CSE, NMAMIT, Nitte, Udupi District, Karnataka, India[1]

Assistant Professor, Department of CSE, NMAMIT, Nitte, Udupi District, Karnataka, India [2]

ABSTRACT: Online reviews for products in many e-commerce sites plays very important role for a both customer in deciding whether to purchase the product or not and the manufacturer to keep track of customer's opinion about the product to improve their sales. But it is not easy to read all thousands of reviews by customers and manufacturers. Opinion mining is used to analyse these online customer reviews. In this paper we made use of document oriented database to store reviews and then sentiwordnet is used for classification of product reviews into positive and negative sentence based on features.

KEYWORDS: Document oriented database; Aspect Extraction; sentiment classification; Sentiwordnet

I. INTRODUCTION

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 just by reading few reviews will not be enough to finalize the products. So it will be helpful for both customer and producer, if the product reviews are mined and presented in summarized manner.

Sentiment analysis or Opinion mining is a part of Natural Language Processing (NLP) which used analyse the opinions expressed by the different users. Sentiment analysis can be performed in three different levels that is at the document level, sentence level and aspect level. In document level, the entire document of particular topic is classified into positive or negative. In sentence level, sentence is classified into positive or negative. In aspect level, the sentence is classified as positive or negative based on aspects or features of that sentence. The feature level sentiment analysis has many challenges and hence it is the current area of interest for the researchers. This paper categorizes opinions using feature level sentiment analysis.

In this paper, we make use of special type of database to store online reviews i.e. MongoDB, one of the NoSQL (Not Only SQL) database. NoSQL databases are used to handle huge volume of unstructured and semi structured data on the web. NoSQL databases are scalable and schemaless. It doesn't store data in the tables with fixed schema like relational databases. These huge amounts of unstructured data and semi structured data cannot be handled by traditional relational databases so we make use of one of the NoSQL database. There are different types of NoSQL database they are key-value database, document-oriented database, column-oriented database, graph database. Here we are making use of MongoDB, which is a document-oriented database that consists of collections and documents. Each collection will have multiple documents and a database have multiple collections. Here values are stored in form of documents; each document contains a unique key with the corresponding value. Using Mongodb we store the real time reviews collected from different websites. Then by applying opinion mining approach we classify and summarize the product reviews, which will help the customers to decide whether to buy the product or not.

Related work is presented in Section 2. Proposed system is discussed in Section 3. Experimental Setup is discussed in Section 4. Results are discussed in section 5. In Section 6 Conclusion is highlighted.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

II. RELATED WORK

This section presents various work related to the opinion mining or sentiment analysis. In [1], the authors mined the customer reviews and summarized the review of products. It focused on mining product features. The authors in this paper have not considered pronoun resolution for opinion sentence. For instance, "*it is quiet but powerful*". Pronoun resolution should to be performed to see what *it* indicates.

The analysis of online product reviews is done using different document level and sentence level, but now ongoing research is on phrase level. In [2], authors makes use of frequent item set mining for extracting the aspects of the product and identifies sentiment orientation of each aspect by supervised term counting based approach. The authors suggested that in future summarization of the extracted aspects need to be done which helps to analyse customers interesting aspects on products.

Mining unstructured and ungrammatical reviews are discussed in research paper [3]. In this work, the authors used support vector machine (SVM) for review classification and to conclude the summary of user's opinion about the product.

Sentiment analysis aims to provide subjectivity and polarity detection [4]. In this paper the author has identified the features from the product reviews using supervised learning method Naive Bayesian algorithm and also determines positive, negative and neutral orientation. The above mentioned sentiment analysis techniques discusses the classification and summarization of customer reviews.

As the growth of data on the Internet, the improvement of corpus and public opinion mining systems need the big data processing techniques to complete tasks. NoSQL database are compared with SQL databases [5], [6]. Here author has discussed that SQL databases are reliable, flexible, robust and scalable. But SQL databases cannot handle unstructured and semi structured data that comes from different web applications. MongoDB is one of the NoSQL database mainly used as a database for Big data processing. Comparison between SQL and NoSQL is done using case study of news website Slashdot [7] which is semi structured d data. In the proposed system we have considered MongoDB as a review database for processing customer reviews. Customer review data is in the unstructured format. In the proposed system the extracted reviews of each product is stored in the form of document. It helps to process review in flexible way by utilizing the features of NoSQL. This system classifies reviews as positive and negative reviews and analyzes customer's interesting aspects on products from the extracted review.

Mining online customer reviews using Aspect level mining [9]. In this paper the author has identified the product features using maximum entropy approach. Here the author has trained a maximum entropy model which as annotated corpus to extract the features of product from the customer reviews.

Mining online product reviews by only considering opinion sentences [10]. In this paper the author for feature extraction consider only those opinion sentences in which the customer has expressed their positive and negative opinions. Here the features are extracted based on probability based algorithm.

III. PROPOSED SYSTEM

The architecture of the proposed system is shown in figure 1. In this proposed system, customer reviews of product collected from different websites and stored into review database called MongoDB. Here the reviews are stored in form of documents. And these documents are given as input for preprocessing stages. Preprocessing includes Stop word Removal and POS tagging. Sentence orientation is used to identify whether it is positive or negative opinion sentence based on aspects. This is done using Sentiwordnet.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

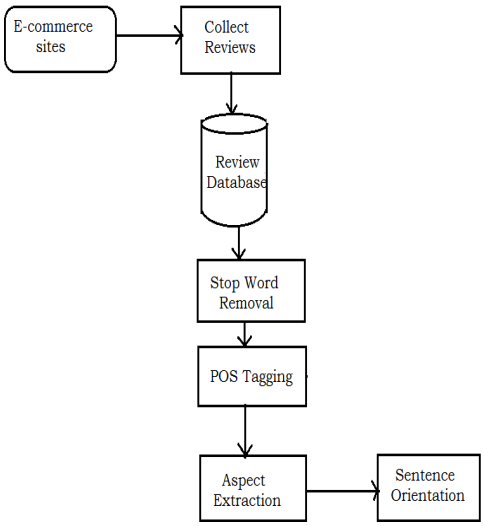

Figure 1: Proposed System

Stop Word Removal

Stop words are unwanted words which are not needed in mining. Example for stop words are 'a', 'an', 'it', 'the' and so on. Reviews which are stored in MongoDB are given to stop word removal. This stop words are removed 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

POS tagger is part-of-speech tagging, which is used to tag each and every word in the given review sentence to its proper part of speech i.e. it tags words as noun, adjective, verb, adverb, conjunction, preposition and interjection. POS tagging of words is necessary to identify features and opinions expressed by the customer about the product in the review sentences. Features 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.

The actual sentence before POS tagging was "This is a awesome product". After POS tagging the sentence using Stanford POS tagger

This_DT is_VBZ a_DT awesome_JJ product_NN

Extracting features of product

In Aspect (or feature) extraction we are going to extract the features of the product. Since we are mining mobile products reviews, some of the features of mobile are size, battery, screen, display, camera etc. These features are noun and noun phrases. Here features are extracted using Apriori algorithm. Apriori algorithm is used to generate frequent item sets. Frequent features generated are to be product features. Here we consider only those item set as frequent which appears more or within 1% minimum support count. The sentence that contain feature and opinion words are nothing but opinion sentence.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

```
for each product review which is POS tagged
        if it contains nouns and adjective words
                extract noun and noun phrases store in
separate text file
                extract adjective as opinion words and its
                nearby adjective as its effective opinion
        end if
        else
                remove the word
        set minimum support count
        count noun words in text file
        if(noun word >= minimum support count)
                extract noun word as feature
        end if
end for
```

Figure 2: Feature and opinion words extraction algorithm

Opinion words and features of products are extracted as shown in figure 2. Here is an example, "The display is awesome", awesome is the opinion word which is effective in the review sentence, to predict the polarity of opinion sentences this effective opinion words are important.

Semantic orientation of opinion words

Words like excellent, awesome etc show positive orientation, while words like poor, disappointing etc show negative orientation. In this paper, we are categorizing only positive and negative polarity using Sentiwordnet. The sentiwordnet is used for assigning three scores such as positive, negative and neutral to each synset of wordnet. The steps are given below.

1. Each sentence which has opinion words is considered. The sentence with extracted feature, each and every opinion words are given values by sentiwordnet.
2. To classify of resulted sentence following two rules are applied:
   a) If sentence has Negation Negative then it is treated as Positive
   b) If sentence has Negation Positive then it is treated as Negative
3. Count the rate of each sentence we are taking one threshold value to categorize positive and negative opinion.
   a) If the rate is above threshold value the sentence is treated as positive opinion.
   b) If the rate is below threshold value the sentence is treated as negative opinion.

Finally the number of positive and negative opinion of each extracted feature in customer reviews is identified. It also gives the extracted feature count of each product that helps to analyse customers interesting aspects on products.

IV. EXPERIMENTAL SETUP

We are extracting real time mobile product reviews from website www.amazon.com using web-crawler. And we stored the extracted reviews into MongoDB. After extracting reviews and storing into MongoDB, we removed the stop words from the review sentences. Then the reviews are spitted into sentences.

Next step is tagging the each and every word in the sentence to their respective parts of speech. This POS tagging is done using Stanford POS Tagger. POS tagging is necessary to identify the noun and adjectives. Nouns and adjectives

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

are extracted which are nothing but features and opinion words. The review sentences based on aspect are classified into positive or negative sentence using the procedure mentioned in the proposed system.

V. RESULTS AND DISCUSSIONS

The database created, named as products. Within product database, collection is created and named as docs shown in figure 3. Collections are nothing but Tables in relational database i.e. in SQL databases. Here the reviews are stored in form of documents, documents are nothing but rows in SQL database.

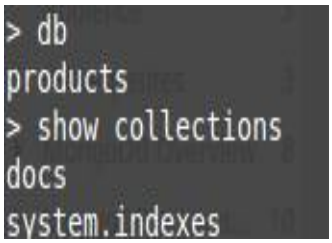

Figure 3: Collections in database products

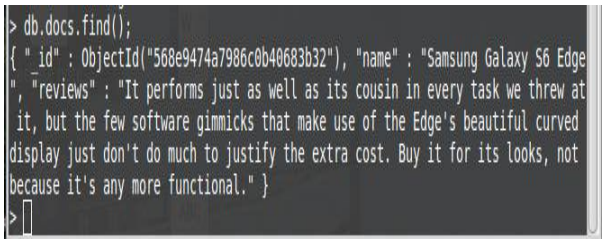

Figure 4: Review stored in MongoDB

The figure 4 shows the reviews collected for Samsung galaxy s6 edge and stored in MongoDB. The name and reviews are fields in the collection docs (which the name of the collection here).

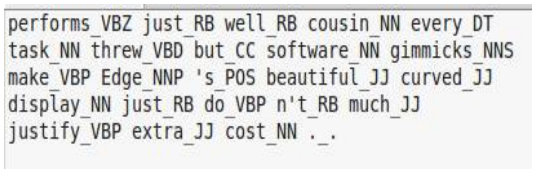

Figure 5: POS tagged words

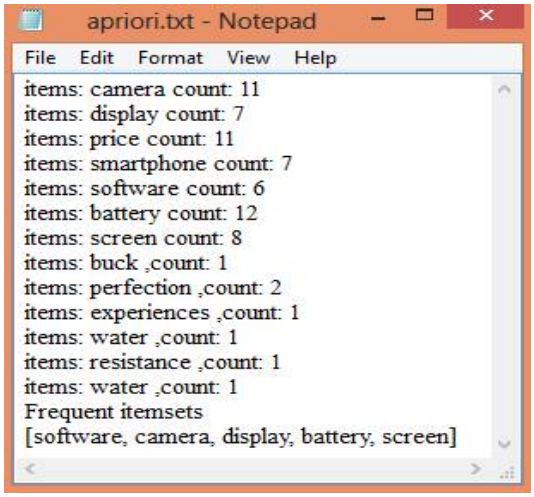

Figure 6: Frequent item set generated

Figure 5 shows the POS tagged sentences. Before POS tagging the review is given to stop word removal to remove all stop word from the review sentences. Then the sentences will be given for POS tagging. From which the nouns and adjectives are extracted which are nothing but features and opinion words. The figure 6 shows the frequent items generated by the apriori algorithm. The item which is more than the minimum support count is selected as feature. The review sentences based on aspect are classified into positive or negative sentence along with feature count of each product using Naïve Bayesian classification technique. The summary is shown in Figure 7.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

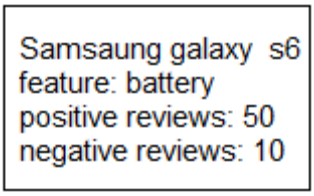

Figure 7: Summary of Aspect

VI. CONCLUSION

In this paper, we collected customer reviews for five different mobile products. Collected reviews are pre-processed using stop word removal and POS tagging. Features are extracted from the review sentences from generating frequent item set using apriroi algorithm. Finally overall summery of a product is given based on the features. Which will be helpful for the purchasing customers to decide whether to buy the product or not.

REFERENCES

[1] M. Hu and B. Liu., "Mining Opinion Features in Customer Reviews," In Proceeding of National Conference on Artificial Intelligence, pp.755-760, 2004.
[2] Jeyapriya A., "Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm", In Proceeding of International Conference on Electronic and Communication Systems, pp. 548-552, 2015.
[3] Taysir Hassan A. Soliman, Mostafa A Esmarsry ,"Utilizing Support Vector Machine in Mining Online Customer Reviews", ICCTA, pp.192-196, 2012.
[4] Nithya R., "Sentiment Analysis on Unstructured Review", In Proceeding of 2014 International Conference on Intelligence and Computing Application, pp.367-371, 2014.
[5] Nishthe Jatana, Sahil Puri, Mehak Ahuja, Ishita Kathuria, Dishant Gosian: "An Survey and Comparison of Relational and Non-Relational Database", International Journal of Engineering Research & Technology, pp.1-6, 2015.
[6] Cornelia Gy_rodi, Robert Gy_rodi, George Pecherle, Andrada Olah, "A Comparative Study: MongoDB vs MySQL", In Proceeding of International Conference on Engineering of Modern Electric Systems, pp.1-6, 2015.
[7] Karamjit Kaur, Rinkle Rani, "Modeling and Querying Data in NoSQL Databases". In: IEEE International Conference on Big Data , pp.1-7, 2013.
[8] Ding, Xiaowen, Bing Liu and Philip S. Yu, '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, 2008.
[9] Gamgarn Somprasertsri, Pattarachai Lalitrojwong. "A Maximum entropy model for product feature extraction in online customer reviews", IEEE International Conference on Cybernetics and Intelligent Systems, pp.575-580, 2008.
[10] Weishu Hu, Zhiguo Gong, Jingzhi Guo. "Mining product features from online reviews". IEEE International Conference on E-Business Engineering, pp.24-29, 2010.