Sentimental analysis of Amazon reviews using naïve bayes on laptop products with MongoDB and R

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042090 doi:10.1088/1757-899X/263/4/042090

Sentimental analysis of Amazon reviews using naïve bayes on laptop products with MongoDB and R

Mohan Kamal Hassan, Sana Prasanth Shakthi and R Sasikala
School of Computer Science and Engineering, VIT University, Vellore-632014, India

E-mail : sasikala.ra@vit.ac.in

Abstract. Start In Today's era the e-commerce is developing rapidly these years, buying products on-line has become more and more fashionable owing to its variety of options, low cost value (high discounts) and quick supply systems, so abundant folks intend to do online shopping. In the meantime the standard and delivery of merchandise is uneven, fake branded products are delivered. We use product users review comments about product and review about retailers from Amazon as data set and classify review text by subjectivity/objectivity and negative/positive attitude of buyer. Such reviews are helpful to some extent, promising both the shoppers and products makers. This paper presents an empirical study of efficacy of classifying product review by tagging the keyword. In the present study, we tend to analyse the fundamentals of determining, positive and negative approach towards the product. Thus we hereby propose completely different approaches by removing the unstructured data and then classifying comments employing Naive Bayes algorithm.

1. Introduction

Smart phones, Laptops and internet have made online shopping very easy. India's internet user base 355 million, registers 18% growth in first 7 months of 2012 IAMAI (Internet and Mobile Association of India.) report. The basic approach can be done protecting the data thus as time was being advanced thus the through the advancement. The technology was also being at rapid change from years to years. Thus the user are being from past century are being getting used to pay high amount to the seller because of additional cost. Such that as the e-commerce industry arrived and established in India in 2005. Thus the era of ecommerce has been started from the year 2005, as the first company Flipkart came to the existence in India by the two People who were working in the Amazon Company from past years as book seller, it determines there mind to have startup in Indian to have an e commerce industry in India in 2005 by bansals. The ecommerce company like myntra, e-bay, snapdeal and paytm came to existence in the Indian. After 2007, such that the customers were first were worried about the service and the quality of the product which was being a major concern and major problem over the e-commerce industry[1].

Thus as the E-commerce industry has started the service and the quality of the product as the user expectation the user being able to trust the e-commerce site to buy the products online in the e-commerce website. Thus there were several techniques adopted by the e-commerce industry to get the belief of the customer by providing the "cash on delivery "and the ratings were provide by the e-commerce industry about the product such that the user can be able to suggest his colleges, friends and relatives about the product review and other details about the product that is available in the e-commerce website[2].

As the time period were started for the e-commerce industry several other company of e-commerce industry are also came in the existence of the market ,which was an competition among each other to provide best ,quality and good product and services to the customer. Thus the e-commerce industry started the process of giving the discount t=for the product which are online and also to make the sakes order by the discount sale and end of season sale criteria were adopted by the e-commerce industry to get through the particular things regard the competition in the market.

Thus we have seen that as the wee are concern about the electronic gadgets we are able to see the following components and their features were given several comments by the customer in the purchase of the laptop product online in the e-commerce website such that the customer and the organizing company of e-commerce website has to understand about the review given the customer for the particular product such that the approach of naive Bayes method has been implemented to identify the review for the particular product is suggested as positive, negative and neutral review by the customer knowledge.

As we know that the customer review are major concern for the industry person because if the reviews are not good and the comment are being in a position that it determines customer to flew away to other e-commerce website to search for the particular product. Such that it will be great loss for the company will be losing his customer based in the review such that the e-commerce industry able to provide good quality product about and the services such that the customer belief of buying the [product is suggested to other, and the organisation of the e-commerce will be satisfying the customer expectation to buy the particular product online in e-commerce website[3,4].

2. Background Study

Hui Song et al. in their paper "Semantic Analysis and Implicit Target Extraction of Comments from E-commerce Websites" Traditional approach always focuses on clear or detailed featured as compared to implicit ones. Thus we have seen that as the wee are concern about the electronic gadgets we are able to see the following components and their features were given several comments by the customer in the purchase of the laptop product online in the e-commerce website such that the customer and the organizing company of e-commerce website has to understand about the review given the customer for the particular product[5].

Yadav, M. P., et al in their paper title "Mining the customer behavior using web usage mining in e-commerce" they explained customer behavior for E-commerce companies using K Mean. With the drastic growth of WWW users can easily find, extract, filter and evaluated whatever they want. With the advancement in technology servers instead to choose from a superstore. Thus there were several techniques adopted by the e-commerce industry to get the belief of the customer by providing the "cash on delivery" and the ratings were provide by the e-commerce industry about the product such that the user can be able to suggest his colleges, friends and relatives about the product review and other details about the product that is available in the e-commerce website[6].

Prashast Kumar Singh research title "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites" there research focus to collect information about what users think about that product and on the basis of it analysis has been done. On the basis of it geographical data can be collected and reviews can be fetched from various sources. In this approach internet slang language and phrases which has helped to gather millions of reviews on social networking sites[7].

Ahmad Tasnim Siddiqui et al. in their paper title "Web Mining Techniques in E-Commerce Applications" explained purchase has been increased as compared to window shopping as it provides millions of ranges. As, companies are able to attract most of the customers. The basic approach can be done protecting the data thus as time was being advanced thus the through the advancement. The technology was also being at rapid change from years to years. Thus the user are being from past century are being getting used to pay high amount to the seller because of additional cost[8].

Songbo Tan et al. in their paper title "Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis" explained in the community of sentiment analysis, such that the determines about the

particular condition of the classifier about which the classifiers are being determined in order to find out the classification of the particular segment to find the token word from the given review by the customer.the e-commerce industry able to provide good quality product about and the services such that the customer belief of buying the [product is suggested to other, and the organisation of the e-commerce will be satisfying the customer expectation to buy the particular product online in e-commerce website.

3. Sentiment Analysis

The sentiment is a natural process of conveying the form of opinion by the customer for the particular product that is available in the e-commerce website .as we are concern about the sentiment approach of the electronic gadgets that is laptop which is being considered to one part of major desks in the Indian w-commerce website there are several recommendation and approaches proposed by the customer regarding the services and the quality of the laptop functionality and it feature and the type of quality of product to customer get the service s from the e-commerce industry through the ecommerce website.

The sentiment analyses is the sentiment of an person not the review of the particular product such that the reviews are contain the word that wanted form of words such that the organization are not able to in determining about the product details and items review whether there review stead is positive review negative review or the neutral review which is enduring to buy for the satisfaction from the particular website. Thus the sentiment plays an important role for the organization which are being introducing several product in their ecommerce website and given way of communicating about the product feature equality and the details. Such that to sentiment analysis came into future and there are several un usual words which does not contain any meaning such that the analysis of the review is done by the sentiment approach of using the navies Bayes classifier considered to be mine of the modern and best method to carry out the sentiment analysis is for any products reviews such that the user an list out about the product whether the particular product to be bought or not.

Such that sentiment analysis [plays a major and key role for both customer and there n=mainly for the organization because it can select their progress in the market due to the competition such that user can flew to other e-commerce website to buy here same product by seeing the sentiment of the product for the same e-commerce product reviews such that the e-commerce company be should be aware of the quality and the review for their product that they are recovering to the customer. And the review s playa san major role as it suggests the customer to find the product to positive, negative one.

4. Mongodb Database

Mongo Db. is the database technologies which is a document oriented database where the data is stored in the form of json file where the data is stored in the key value pair ,such that mongo dB as the CRUD operation[9] through which we can manage the operations in the mongo dB database which is considered to a database where we can be able to add the particular connection any technologies to have the connectivity with the particular front end such that we can interact which the data can access and manage.

The package we are going to perform sentiment analysis with Mongo dB using rmongodb package.

```
>library(rmongodb)
> library(rjson)
```

Another library rjson is used, which converts r objects into json objects and vice versa.

5. Navies Bayes Classifier

It is a method to approach the following segments of the review which the customer are providing such that the review are consisting of several un wanted words such that classifier are being in search of the tag words such that the defined tag words in the vectors which are positive and negative such that the there are several condition to calculate the following probability of calculating the points to

determines about the classification of the major review where there are several special symbol which must be removed by the classifier[10,11].

Thus the navies' Bayes classifier is an method and an approach of statistics to get through the vectors contain the word and several library package to determine that the particular tag word must be counted in order to give there result about an accuracy of data to which we can determine whether the user can be able to identify the review rating about user reviews[16,17,18].

$$p(c/d) = \frac{p(c)\,p(d/c)}{P(d)}$$

5.1 Evaluation Setup

The evaluation step of the sentiment analysis classifier is based on the approach of the ***Text Pre-Processing***

The text pre-processing is the process of refining the review which is in the form of text where several unwanted words are being there in the review part of text such that is provide a not meaning full sentences.

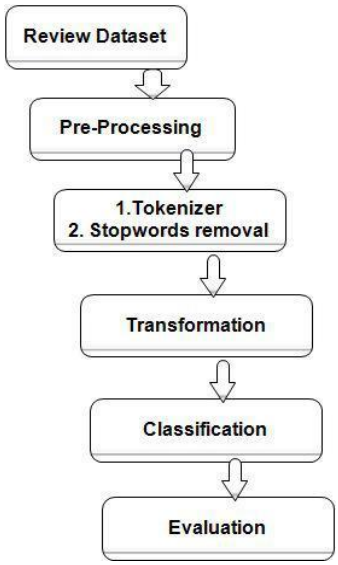

Figure 1. Steps to Evaluate Sentiment Analysis

The words to be removed from the review part are by the part of the tagged words from the word which are considered to be stop words. Like there are several words that is and ,or and is etc. doses no provide an legitimate meaning for the review to be understand. The next process is to transform the word such that the transformed word can be used to identify the meaning of the review accurately, such that the decision can be made through the list of words. The final step is to determine the considered features that are to be selected in order to determine about the particular ways of identifying the approach of understanding the review of the customer[12,13].

6. Implementation of Algorithm

Algorithm step 1:

Algorithm that is being implemented for getting the review

Require: The Products Reviews

Ensuring: The sentiments of User comment.

1. Fetching all comment.
2. Conversions of the unwanted format of comment data to a wanted understandable format document.
3. The process of determining the tokens that is to be as keyword.
4. The implementation and processing of unwanted stop character and the required tagging of word by the determined tagged word.
5. The search for the particular word that is the correct word.
6. Apply Nave Bayes classifier.
7. Compute sentiments using Algorithm 2
8. The last step is the analysis of reviews, through the sentiment analysis and returning of the score for each review.

Algorithm 2:
Algorithm to calculate the review orientation

```
1. Procedure Review Sen( )
2.      start
3.      for the determined each review sentence seni
4.      start
5.      sen = 0;
6.      For the each review belong to word r in seni
7.      sen + = Word Sense (r, seni);
8.      /* Pos =1 , Neg =-1*/
9.      if (sen >0) seni' s sen= Pos;
10.     else if (sen<0) seni' s sen = Neg
11.     end for;
12.     stop
```

Algorithm 3:

```
1.      Procedure Word Sense (wrd, sent)
2.      start
3.      sen = introduction of the determined words ;
4.      if(there are  NEG_WRD appeared close and around to the words in sent)
5.      sen = opposite(sent);
6.      stop
```

7. Results and Discussion

To perform sentiment analysis using MongoDB first we need to establish a connection between MongoDB and R language. For that we need to run Mongodb.exe and Mongo.exe applications after that we need to import json file from MongoDB. After the zip files are extracted now we can go to the R console and run the following command in order to connect the R and MongoDB.

```
>library(rmongodb)
>m<-mongo.create()
>ns<"DatabaseName.CollectionName"
example: (DbName:sales,CollectionName:product)
```

```
n<-mongo.create()
ns<-"database.collection "
mongo.is.connected(n)
[1]=TRUE
ns<-"Services.Rating"
mongo.is.connected(n)
[1] TRUE
```

Figure 2. Connection Establishment between MongoDB and R

Data set taken can be done categorization process indicating the positive-ness or negative-ness of a given sentence. For this to carry out we will use Naive Bayes algorithm[14,15], its general. The implementation of the sentiment analysis for the review classification for the review given by the user is based on the match of the word which is determined in the keyword as determined in the R code for the implementation. Word tokens for positive and word token for negative, which can identify either the review is positive or negative throughout the dataset. As the determined tokens and the sentiment for the data and the information that is obtained from the analysis and implementation for the particular approach from the original dataset.

Table 1. Sentiment Scores Extracted from Dataset

S.No	Score
1	1
2	-1
3	2
4	0
5	0
6	-1
7	0
8	0
9	1
10	1
11	1
12	-1
13	-1
14	1
15	2
16	2
17	1

Here, Navies Bayes classification to construct the matrix.

The matrix for the taken dataset results in:

Table 2. Sentiment Classification

Positive	Negative	Positive/Negative	Sentiment
9.4754700	0.44545322	21.2715265477714	positive
16.533678	9.47548756	1.744559248756	positive
8.75325476	0.45714567	19.71548736845	positive

1.03457468	8.78322547	0.1174265478734	negative
9.4754700	0.44545322	21.2714875434157	positive
1.03127457	9.47547243	0.1024578943567	negative
8.78234715	0.4454532	19.715477234057	positive
9.47156731	8.78232285	19.715477234057	neutral
8.78234756	0.44545322	37.115834295331	positive
8.79548721	18.5054868	0.4745772340541	positive
25.6707451	26.8423546	0.9563507468519	positive
17.2277489	0.44547168	38.671883607064	negative
25.4784361	0.44547124	57.624719853427	Positive
9.47547812	9.4754700	1	neutral

The results had shown sentiment classification and sentiment scores of the positive and negative of different reviews by plotting graphs. These graphs are eventually the evidence for representing the following presentation of the following criteria, thus the following graph determines that are the sentiment score for the following reviews for the given reviews are given as below which ranges from -3 to +2 as it specific the particular word substituting among the reviews. The second graph provides the relationship between positive and negative reviews for the product.

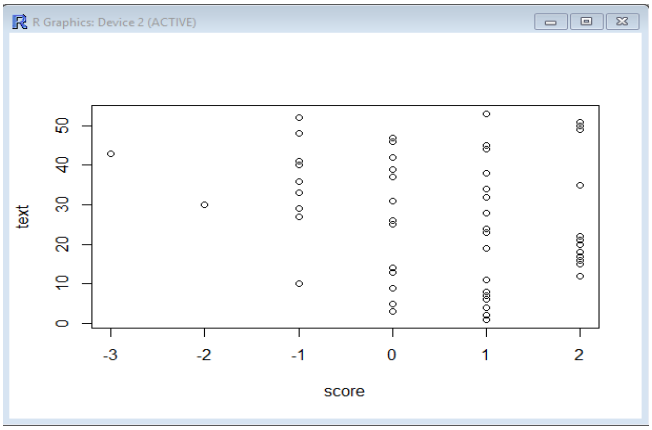

Figure 3. Sentiment Score Information for Word Token

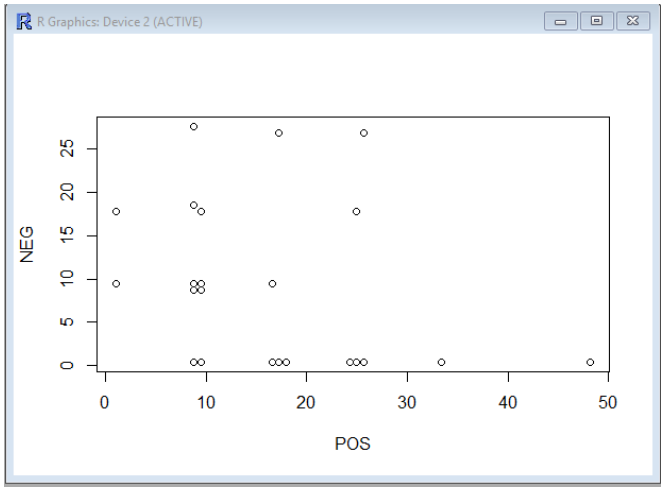

Figure 4. Navies Bayes Sentiment Classification

7.1 Comparison

Comparing different classification techniques and evaluating their accuracy for the sentiment analysis.

Machine learning approach: It mainly depends upon the machine learning algorithms and mainly evaluate for sentiment analysis of document analysis. It accuracy can be approximately 71.7% for 3 categories and 46.9% for 5 categories. Main disadvantage is that it can be only applied to the reviews that are written in English as it takes words from its database.

Bayesian Networks (BN): Naive Bayes classification has independence among its features while Bayesian networks can be said that it has dependence for all features. It can be used as acyclic graph and features as nodes and has various relationships between them. Accuracy for the probabilistic model is approximately 73%. BN's can be difficult to perform for the unsupervised models as the correlation for the same clusters and actual features is not the same.

As the Naïve Bayes and Bayesian networks can show performance approximately same[13]. It solves the general problem that occurs between positive and negative classification. Generally NB classifier has a way to increase. But, this will become obstacle when calculating the average of two classes it decreases average accuracy value. Kang and Yoo showed in their paper that their accuracy is improved. It shown that it is better than NB and SVM.

Accuracy can be calculated based on the paper proposed by A. Hamouda and M. Rohaim [12]. In that paper all their results of restaurants reviews are extracted and it contains 2000 samples. And these samples include 1000 positive and 1000 negative reviews. Based on these we are going to perform Naive Bayes accuracy for 20000 samples of reviews which includes positive, negative and neutral. Reviews include parameters like single words, stop words Removal, bigrams, Bigrams and stop words.

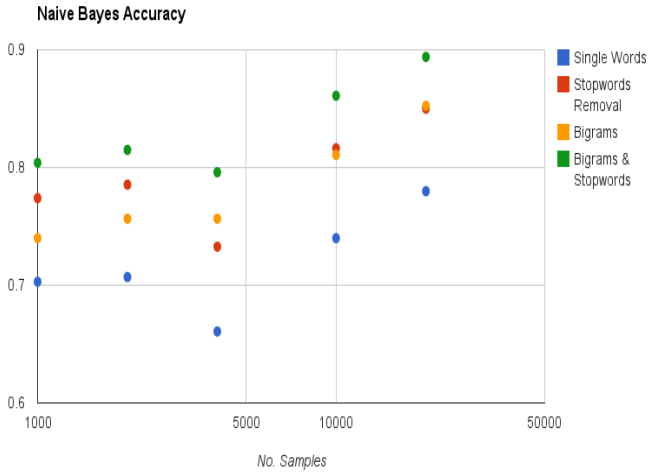

Figure 5. Naïve Bayes Accuracy

8. Conclusion

Instead of some thousand products in a superstore, consumers may choose among millions of products in an online store to satisfy the personalization demands. In this paper, we use Naïve Bayes algorithm and semantic decision tree to classify the polarity of comments given on e-commerce websites. First, we use a web crawler to fetch comment on a particular web page. The spelling correction is done to make the most sensible comment for knowing the polarity of words using Word Net dictionary. Then stemming is performed to remove the stop words. After classifying the positive and negative words using Naïve Bayes algorithm, the overall polarity is calculated using decision tree. In future it can be extended our study on framework developed websites where tags are hidden in browser and we will add prevision of adding other languages words in dataset for more accurate results.

References

[1] ChandraKala S and Sindhu C 2012 Opinion mining andsentiment classification: a survey *ICTACT J. Soft Comput.* pp420-427.

[2] Kim, S.M. and Hovy, E. 2004 Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics* pp. 1367.

[3] Liu B 2010 Sentiment analysis and subjectivity In: Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca pp. 627-666

[4] Liu B, Hu M and Cheng J 2005 Opinion observer: Analyzing and comparing opinions on the web In Proceedings of the 14th International Conference on World Wide Web, WWW '05, pp 342–351.

[5] Song, H., Chu, J., Hu, Y. and Liu, X., 2013, December. Semantic Analysis and Implicit Target Extraction of Comments from E-Commerce Websites. In *Software Engineering (WCSE), 2013 Fourth World Congress on* (pp. 331-335). IEEE.

[6] Yadav, M.P., Feeroz, M. and Yadav, V.K., 2012, July. Mining the customer behavior using web usage mining in e-commerce. In *Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on* (pp. 1-5). IEEE.

[7] Singh, P.K., Sachdeva, A., Mahajan, D., Pande, N. and Sharma, A., 2014, September. An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites. In *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-* (pp. 329-335). IEEE.

[8] Siddiqui, A.T. and Aljahdali, S., 2013. Web mining techniques in e-commerce applications. *arXiv preprint arXiv:1311.7388*.

[9] Sasikala, 2016. Research based literature survey and analysis on various sharding techniques. *IIOAB JOURNAL*, *7*(9), pp.479-494.

[10] Tan, S., Cheng, X., Wang, Y. and Xu, H., 2009. Adapting naive bayes to domain adaptation for sentiment analysis. *Advances in Information Retrieval*, pp.337-349.

[11] Pang B, Lee L 2004 A sentimental education: Sentiment analysis using subjectivitysummarization based on minimum cuts In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04.Association for Computational Linguistics,Stroudsburg, PA, USA pp 271

[12] Stanford 2014 Sentiment 140. http://www.sentiment140.com/

[13] Sentiment Analysis in R http://andybromberg.com/sentiment-analysis/.

[14] Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, *2*(1–2), pp.1-135..

[15] Liu B 2012 Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan &Claypool Publishers.

[16] Hu M, Liu B 2004 Mining and summarizing customer reviews In Proceedings of the tent ACM SIGKDD international conference on Knowledge discovery and data mining, pp.168–177

[17] Hamouda, A. and Rohaim, M., 2011, January. Reviews classification using sentiwordnet lexicon. In World congress on computer science and information technology. IAENG..

[18] Hanhoon Kang, Seong Joon Yoo, Dongil Han 2012 Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews *Expert Syst Appl*, 39 pp. 6000–6010