Challenges for Information Retrieval in Big data: Product Review Context

Sanjib Kumar Sahu
Dept of Computer Science,
Utkal University, Odisha, India

D. P. Mahapatra
Dept of Computer Science,
NIT, Rourkela, Odisha,

R. C. Balabantaray
Dept of Computer Science,
IIIT Bhubaneswar, Odisha,

ABSTRACT
The ever increasing scale of e-commerce has today presented a big range of choice for the customer. Customer uses online product reviews as a primary criterion to make a decision for his purchase. These product reviews are scattered all around the internet, and this data has a great potential value. However, it is also unstructured and written in a natural language, which poses great problems for data mining and data analytics. The scale, non-uniformity and complexity of product reviews make them classic big data elements. This paper discusses the big data challenges and opportunities involved in mining and analytics of product review data. It formally studies the problem under a big data framework and formulates a plan for the extraction, mining and analysis. This paper also reviews some of the mining approaches for product reviews and implemented feature/attributes based method for finding the review of products.

Keywords
Big Data; Information Retrieval;Data Mining;Product Reviews; Text Mining; Sentiment Classification; e-commerce;

1. INTRODUCTION
E-commerce is a 21st century innovation which has changed the 150,000 years old practice of trading the physical goods [1]. It has brought a whole marketplace to our phone and laptop screens, with a huge variety of capital goods available to purchase at any time. This is becoming a major driver of economic growth in populous and developing economies like India and China, where booming businesses are bringing investments, creating jobs and most importantly empowering the customer [2]. There has never been another time in history where a customer had so wide range of options. There are so many merchants selling hundreds of brands and thousands of products which can be shipped to every corner of the world, that customers are spoilt for choice today. For example, a simple search for a mobile phone in the price range of Rs. 10,001-18000 on a popular online store *Flipkart* brings more than 300 product results. The dilemma of a customer today is to make a decision which gives him the best value for his money.

Product reviews have become the primary resource for a customer to get all the information about any item and mostly a wise customer decision is a result of reading tons of reviews on the internet. These reviews come from many places. There are specific websites and blogs where professionals use a product and post their impressions, there are many online forums where users of a product can submit their comments, and merchants selling these products also allow customers to rate and review them on the web. A huge amount of this review data is being generated everyday, which is mostly free to access and has a great value to the customer. Yet, procedural analytics for this data have been highly neglected, majorly due to its associated complexity. Such reviews are highly unorganized as they are written in a natural language. They are ever growing and they originate from multiple heterogeneous sources. Moreover, the user perception of a product also changes over time due to the availability of better alternatives in the marketplace. The characteristics of review data fit perfectly to the classic big data definition. This data is voluminous and heterogeneous, it originates from autonomous sources with distributed and decentralized control, and it has complex and evolving nature. Therefore, the techniques and practices traditionally employed for big data must be adopted for analysis of such product reviews.

There are tremendous advantages and opportunities of undertaking such data analysis [3]. It creates a value out of the unstructured data floating around the web. The analysis report has great significance for a customer, who can now focus on the key positive and negative sentiments of a product without wasting a lot of his time searching on the internet, compare the product to other similar products in the market and shortlist his feature preferences to make a quicker and much more effective decision. Such reports are also more beneficial than limited surveys, and manufacturers can use them to understand the market demand better to design improved and more attractive products. The product reviews serve as a valuable feedback which improves the profitability of all the stake-holders involved in an online retail business like the manufacturers, the retailers and the customers.

The roadmap to the remaining part of the paper is as follows. Section-II discusses the big data challenges and solutions for the product review analysis. It also defines and classifies the complexity level of this data. Section-III proposes a general process for the analysis of such data. Section-IV discusses the existing methodologies of product review. Section-V discusses about sample collection, result analysis and procedure to a methodology for product review. Finally, the conclusion and future development for investigation is presented in Section-VI.

2. DATA CHARACTERISTICS
In a recent and widely cited paper [4], Wu et.al have proposed HACE theorem to model big data characteristics. According to HACE theorem, Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. This paper experiments above mentioned model for understanding the characteristics of product review data. Later, this paper will also classify this data using the popular matrices to understand the challenges involved in the analysis. Below, first explain the basic concepts of HACE theorem.

2.1 Huge Heterogeneous Data with Diverse Dimensionality
Product reviews are unstructured and complex. Generally there is no common standard template, and rather these are

written in a natural semantic language. Moreover, there are different methods a user can use to indicate the product impressions. Some reviewers list the pros and cons of a product and some write full opinion paragraphs without explicitly listing their specific comments. Almost all major merchants like *Amazon* and *Flipkart* usually allow reviewers to give star-ratings or thumbs-up to a product. Moreover, a product like a mobile phone has a diverse set of features, and every feature can evoke different opinions. For example, a phone can be light and durable, but it may have a resource-hogging operating system or poor camera. It is important to include this extra dimensionality while analyzing the product reviews since it encourages the customer choices based on their individual preferences and indicates the market requirements for a specific feature. All these issues become major challenges to aggregate the data from different sources.

2.2 Autonomous Sources with Distributed and Decentralized Control

The unorganized product reviews are always generated from autonomous sources, as they are essentially human responses. The human sentiments can vary from person to person, based on the individual requirements, preferences and tastes. Also, a manufactured batch can differ from another batch of products where quality control methods are not stringent. There is no central authority dictating their opinions, and hence there can be a wide range of differences in the beliefs about a particular product across various market segments.

2.3 Complex and Evolving Relationships

No product is bad just by itself since it serves to a requirement. Instead, a product is just better or worse as compared to other alternatives available in the market. The user opinions have a complex and time-evolving relationship which depends on these alternatives. Goods produced in a free market always face competition for customer attention by providing them with a better price-to-value ratio. Manufacturers regularly use latest technologies and economic practices to change the market dynamics into their favor. A product may not be receiving negative reviews initially, but the opinions may change after the introduction of a new alternative product into the segment. On the other hand, a negatively seen product may turn for better after a price adjustment in the market. It is necessary to account for this evolving nature of data in the analysis for a deeper understanding of the market demands.

From this investigation, the establishment of the product reviews is fitted perfectly into the big data model and they must be processed accordingly. The data characteristics can be classified using the traditional big data matrices [5] as follows:

Volume: The amount of the considered data is quite large. Each feature of a product can attract thousands of reviews in reality. The analysis methods must be scalable to the data volume.

Variety: The variety of such data can also be very considerable. Variety may differ for different products, and it must be known to the data analyst. Understanding the variety helps effective use of the data.

Velocity: As compared to other sources such as twitter feeds, which generate huge amount of data every second, the complete product reviews have relatively less generation velocity. This reduces the processing requirements and the analysis can be set up at sparsely scheduled intervals.

Variability: Variability is a huge problem when dealing with unstructured product review data. Effective data mining algorithms must be developed to extract the useful information out of the review text. Some of the product review mining methods are discussed in section IV.

Complexity: The complexity of data can also be significant since it needs to be linked to the alternatives available in the market at a time instant. Efficient methods to manage this complexity must be developed.

3. ANALYSIS PROCEDURE

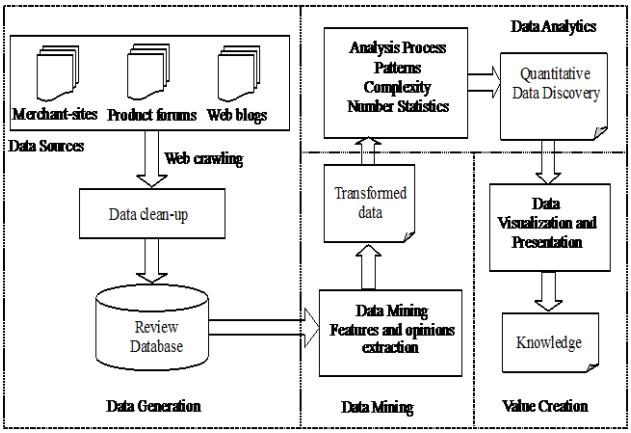

Fig 1: A General Data Analysis Process for Product Reviews Indicating the Primary Activities and Constituents.

A generic process flow which can be followed for the analysis of product review data is shown in Fig. 1. It can be divided into the following major activities. From Fig-1, it is observed that, there are four major activities involved in the analysis of product review data. They are (3.1) Data Generation (3.2) Data Mining (3.3) Data Analytics (3.4) Value Creation. The complete explanation is given below.

3.1 Data Generation

The first step to analyze the data is to actually get the data. Although there are many sources on the internet where the product reviews are available in public domain, they are scattered, unorganized and unformatted. It is very important to extract this data from the web and store them in a central repository for any further operation. For this, the data sources must be identified first. Then, web crawlers can be designed to automate the process of selecting and extracting the required data fields from unstructured HTML pages and store them in formatted files. This data is then required to be cleaned up further before storing them in the review database.

3.2 Data Mining

Data mining is the most crucial aspect of this whole problem. It is also tricky, since the reviews are written in natural language, and one must use semantic approaches to extract the useful information from the reviews in a quantitative manner. There are many such mining approaches, which are discussed in Section IV. Note that data mining is different than text summarizing, since the concept is to extract out the opinions about different features and the complexity between these opinions, in a form which can be easily processed by the analytics methods. Also, not all the product review data are

unstructured. The star-ratings and score grading of a product feature can be used directly for quantitative estimates.

3.3 Data Analytics

Once the data is mined and the important characteristics are extracted, different analytics methods can be implemented to find the inherent relationships among these data. These analytic methods can focus on the pattern discovery, and statistical analysis for complexity and numbers. While data analytics schemes are very important, they are not the most challenging part of the whole process. Such analytics methods have traditionally been used for many of the big data applications, and they should also operate fairly on a well-mined product review data. This exercise results in the discovery of quantitative matrices which are useful for making economical and technological decisions by individuals or firms.

3.4 Value Creation

The data analytics results are of no use to the end decision maker unless they are presented in a concise and efficient form. Data visualization forms a big and important part of this process, as the end value created from the data is as effective as the way data is expressed. One must use the already established business charts and graphing techniques. However, new and efficient visualization methods must also be explored which apply specifically to the operated dataset. An efficient presentation lends itself to the final knowledge, which is the ultimate objective for this whole assignment. It should help in making quick personal and business decisions.

4. EXISTING METHODOLOGIES FOR PRODUCT REVIEW

Many researchers have explored the mining and summarization methods for user reviews. A few noticeable approaches have been discussed in this section. Dave and others in [6] used supervised machine learning methods for semantic classification of reviews. Using available training corpus from some websites, where every review already had a class (e.g., thumbs-up and thumbs-downs, or some other quantitative or binary ratings), they designed and experimented a number of methods for building sentiment classifiers. They have shown that such classifiers perform quite well with test reviews. They also used their classifiers to classify sentences obtained from web search results, which are obtained by a search engine using a product name as the search query. However, the performance was limited because a sentence contains much less information than a review.

In [7], Morinaga *et al.* compared reviews of different products in one category to find the reputation of the target product. However, it does not summarize reviews, and it does not mine product features on which the reviewers have expressed their opinions. Although they do find some frequent phrases indicating reputations, these phrases may not be product features (e.g., "doesn't work", "benchmark result" and "no problem(s)"). In [8], Cardie *et al* have discussed opinion-oriented information extraction. They aim to create summary representations of opinions to perform question answering. They propose to use opinion-oriented "scenario templates" to act as summary representations of the opinions expressed in a document, or a set of documents.

Hu and Liu [9] have followed an interesting approach of feature based classification of user reviews, where they mine individual features of a review by analyzing each sentence and also extract the opinion orientation about that review, whether it is positive or negative. Their method also does not require any training data. Their approach is based on the following three steps.

1. Mine the product features that have been commented on by customers. They used both data mining and natural language processing techniques to perform this task.
2. Identify opinion sentences in each review and decide whether each opinion sentence is positive or negative. Only those lines which contain one or more product features are identified as an opinion sentence. The *opinion orientation* of each sentence (whether the opinion expressed in the sentence is positive or negative) is decided using a set of adjective words (called *opinion words)* which are normally used to express opinions.

This paper proposes to use a feature based review mining method discussed above for this work. This method adapts itself very well to the big data analysis. The study has taken 96 features of the Mobile phones like: screen quality, battery, sound, memory, design, heating,GPRS etc. and these attributes are used for classifying the reviews into positive and negative given by different user on mobile phones.

5. RESULT AND DISCUSSION

In this section the online customer reviews are extracted from Web.

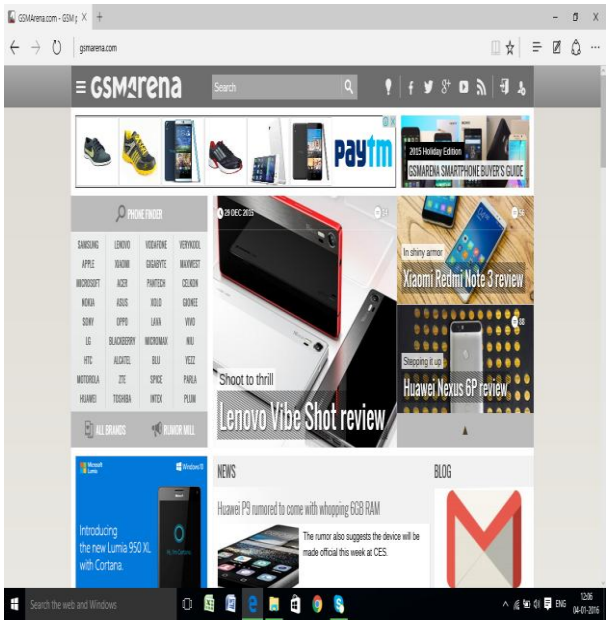

Fig 2: Website Showing Product Review

The above is a sample site which is used to extract the reviews. The study extract the review of different types of mobile phones like Sony, Nokia, Samsung, Vevo, Motorola, HTC, Blackberry, Intex, LG, Lenovo, Huawei, Micromax and Acer etc.

The analysis retrieves information about specific products or features and classifies it as "positive" or "negative". Sometimes it is mixed a review like some positive and some negative. Sometimes it is a review like little bit positive and little bit negative. That's why the feedback related to specific features matters more than overall feedback.

Sample reviews are

- Cell phones with great quality front camera or back camera;
- Good resolution and its screen quality is too good; I am not impressed with this camera;
- I am not impressed with battery back-up; I am impressed with processor speed;
- I have a very hard time getting internet speed as this phone doesn't support 4G or 3G;
- Not impressed at all.

A product review consists of different opinions, the objective is to analyse the expression of view and segmental division into positive or negative. Consider the review "How Can people use NOKIA LUMIA 830, it's battery backup and Touch Screen quality is worst, but the design structure isn't that bad".

In the above review the "Battery back-up" and "Touch Screen Quality" words are used for mobile domain. The words "Worst" or "not that bad" refer to battery back-up and Design-up Structure features respectively. "Worst" is negative opinion and "not that bad" is a positive opinion.

Before starting the result calculation of Precision, Recall,

F-measure, ROC Area based on the review, let's explain the concept of Precision, Recall, F-measure, ROC Area

PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

Precision = (Total number of sites selected for evaluation) /

(Sum of the scores of sites retrieved by search engine)

RECALL is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

Recall= (Total number of sites retrieved by search engine)/ (Sum of sites retrieved by search engine)

Mathematically Precision and Recall are defined as [30]

Precision= (Tp)/ (Tp+Fp)

Recall= (Tp)/ (Tp+Fn)

Where,

Tp-True positive

Fn-False negative

Fp-False Positive

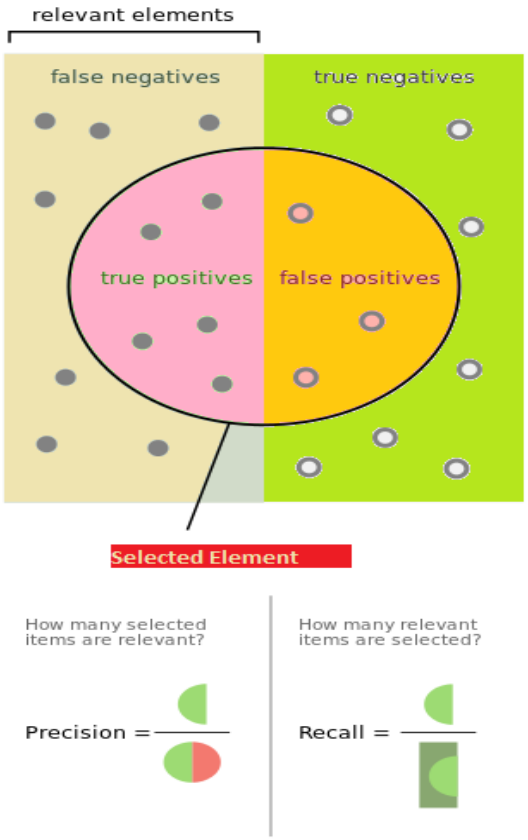

Fig 3: Precision and Recall Description

F-Measure: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score. Mathematically

F=2(Precision x Recall)/ (Precision + Recall)

ROC Curve: Receiver Operator Characteristic (ROC) curves are generally used to present results for binary decision problems in machine learning. However, when dealing with highly skewed datasets, Precision-Recall (PR) curves deliver a more informative picture of an algorithm's performance [30].

This experiment has followed the following steps for getting results:

Collected the reveiw from different mobile websites.

Extracted the features from the reveiws.

Tagged the opinion of the reveiws

(i,e +ve &-ve) by the developed software tool

Does training by taking sample data.

Does testing by taking sample data.

Obtaining the result.

This Experiment have collected 250 samples and later increased to 500 plus samples related to review of mobile phones from the web site mentioned in Fig-2 and the data is stored in a plain text file like in Fig-4. From this review, 36 keywords have been identified and later increased to 96 keywords like "back-up", "Battery", "Touch", "camera", "waterproof", "screen", "processor", "3G", "Memory", "GPRS", "Design" etc displayed in Fig-9. These key words

are the features/attributes of mobile phones. Weka tool (shown in Fig-6), SVM light tool have been used for analysis of result. The feature matrix has been created by the developed software tool (shown in Fig-5),This feature matrix will be input for both Weka as well as SVM. For this data, the 10 fold validations have been conducted with the help of Weka tools and sequence of procedures and the respective snapshots of the processing are given below (Fig-7 to 15).

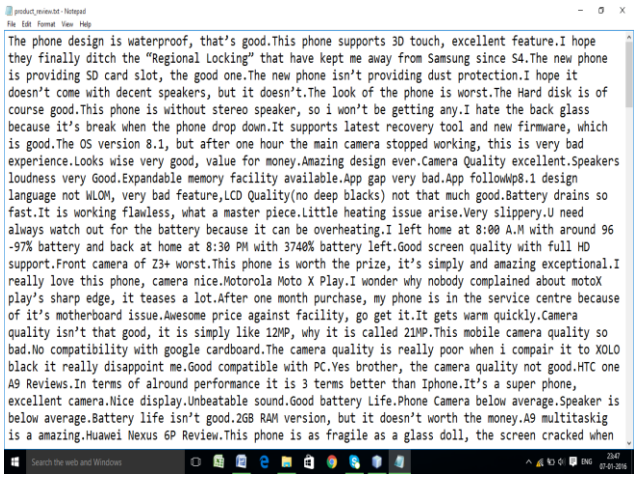

Fig 4: Review Data Collected and Stored in a Plain Text File.

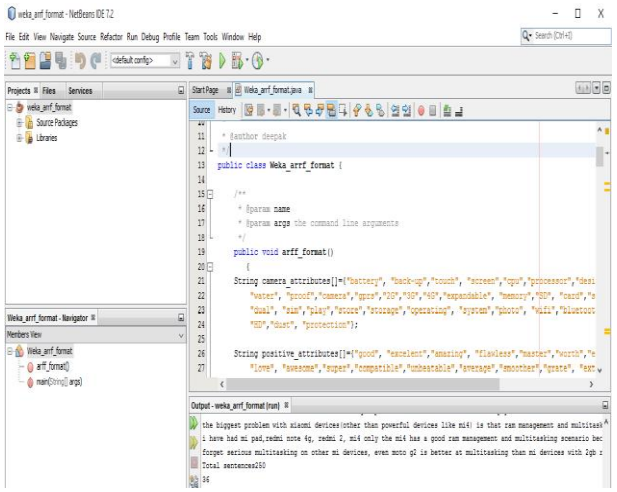

Fig 5: Developed Software Tool Converting Text to Attributes and Matrix Based on Features

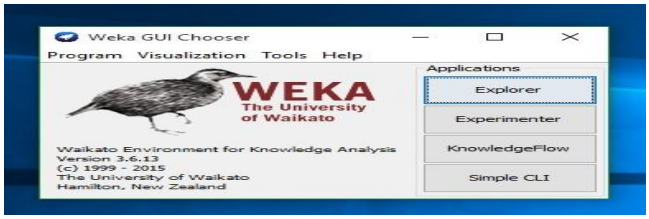

Fig 6: Weka Tool Interface

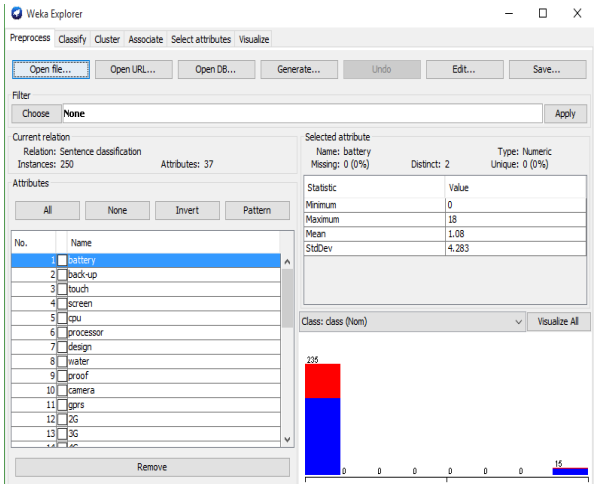

Fig-7: Showing Attributes of Mobile Review for 250 Reviews.

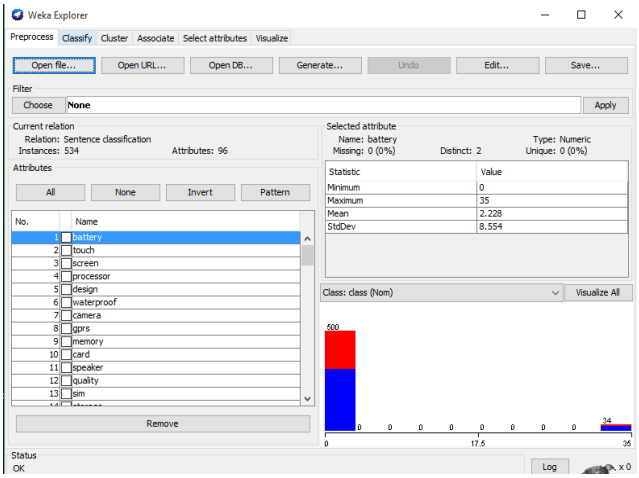

Fig-8: Showing Attributes of Mobile Review for 500 Reviews.

```
@relation 'Sentence classification'
@attribute battery numeric
@attribute back-up numeric
@attribute touch numeric
@attribute screen numeric
@attribute cpu numeric
@attribute processor numeric
@attribute design numeric
@attribute water numeric
@attribute proof numeric
@attribute camera numeric
@attribute gprs numeric
@attribute 2G numeric
@attribute 3G numeric
@attribute 4G numeric
@attribute expandable numeric
@attribute memory numeric
@attribute SD numeric
@attribute card numeric
@attribute speaker numeric
@attribute LCD numeric
@attribute quality numeric
@attribute dual numeric
@attribute sim numeric
@attribute play numeric
```

Fig 9: Showing Attribute Properties in the Input File.

```
0,0,0,0,0,0,0,0,0,24,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,negative
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,positive
0,0,0,0,0,0,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,positive
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,positive
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,positive
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,positive
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,negative
0,0,0,0,0,0,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,negative
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,negative
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,negative
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,negative
```

Fig 10: Showing the Input Matrix for Weka .

For the above matrix data, A 10 fold validations has been conducted and the results are as follows and shown in Fig-(11-12) with the help of Weka tool.

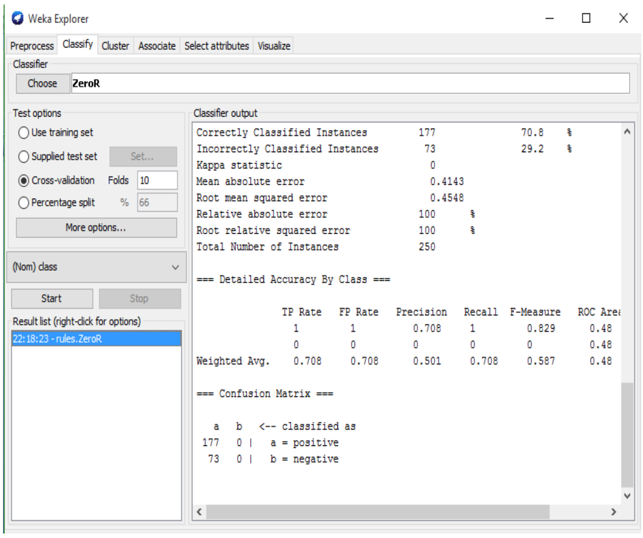

Fig 11: Analysis of Precision, Recall and ROC Area for 250 Mobile Product Review using Weka Tool.

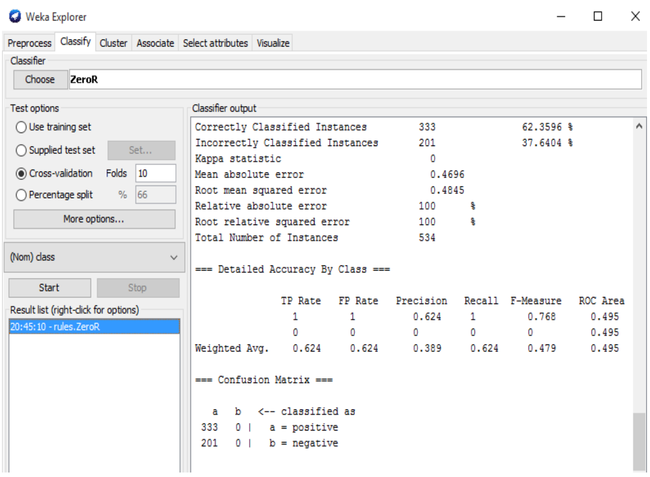

Fig 12: Analysis of Precision, Recall and ROC area for 500 Mobile Product Review using Weka Tool.

We have also analyzed the results with the help of SVM (Support Vector Machine). The model is created out of 100 samples is shown in Fig-13 and subsequently the testing has been done with the help of 20 and 103 samples separately are shown in Fig-(14-15). The output of the model file is as follows.

```
Scanning examples...done
Reading examples into memory...OK. (96 examples read)
Setting default regularization parameter C=0.2586
Optimizing.........................................done. (44
iterations)
Optimization finished (5 misclassified, maxdiff=0.00071).
Runtime in cpu-seconds: 0.01
Number of SV: 60 (including 37 at upper bound)
L1 loss: loss=15.49768
Norm of weight vector: |w|=2.96358
Norm of longest example vector: |x|=3.87298
Estimated VCdim of classifier: VCdim<=80.04505
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=56.25% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>41.30% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>41.30% (rho=
1.00,depth=0)
Number of kernel evaluations: 3505
Writing model file...done
```

Fig 13: Model Generated in SVM Tool.

The output after testing is described in the Fig-(14-15).

```
Reading model...OK. (60 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 80.00% (16 correct, 4 incorrect, 20 total)
Precision/recall on test set: 100.00%/73.33%
```

Fig 14: Analysis of Precision, Recall for Testing 20 Mobile Product Review Using SVM Tool.

```
Reading model...OK. (347 support vectors read)
Classifying test examples..100..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 76.70% (79 correct, 24 incorrect, 103
total)
Precision/recall on test set: 76.70%/100.00%
```

Fig 15: Analysis of Precision, Recall for Testing 100+ Mobile Product Review using SVM Tool.

6. CONCLUSIONS

In this paper, the investigation has been done to study the feasibility of applying data analytics techniques to product reviews to generate greater value out of the unstructured data. This paper has established that this problem falls under the domain of big data analysis because of the associated data volume, non-uniformity and complexity. The product reviews have been classified according to popular big data modeling matrices, and the associated characteristics and challenges have been identified. In this paper a new approach has been proposed a general process flow for a formal analysis of product reviews. The data must be first extracted and mined before applying any analytics. This paper have surveyed some product review data mining methods, and found that the feature based mining generates the quantitative data which can be operated using analytical algorithms. It has been observed that SVM gives better result in terms of precision and recall. Still lot of improvement is required in this domain for further improving the result stated above, i.e. precision recall, etc. The study is in the process of using semantics for identification of features/attributes to improve the results i.e. to achieve ROC area nearly equal to 1.

7. REFERENCES

[1] Watson, Peter (2005). *Ideas: A History of Thought and Invention from Fire to Freud.* New York: Harper Collins Publishers. ISBN 978-0-06-621064-3.

[2] Raven, Peter V., Xiaoqing Huang, and Ben B. Kim. "E-business in developing countries: a comparison of China and India." *International Journal of E-Business Research* 3.1 (2008).

[3] McAuley, Julian, and Jure Leskovec. "Hidden factors and hidden topics: understanding rating dimensions with review text." *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013.

[4] Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on* , vol.26, no.1, pp.97,107, Jan. 2014.

[5] Sagiroglu, S.; Sinanc, D., "Big data: A review," *Collaboration Technologies and Systems (CTS), 2013 International Conference on* , vol., no., pp.42,47, 20-24 May 2013.

[6] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews."*Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.

[7] Morinaga, Satoshi, et al. "Mining product reputations on the web." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[8] Cardie, Claire, et al. "Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering." *New directions in question answering*. 2003.

[9] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews."Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[10] Olson, David L.; and Delen, Dursun (2008); *Advanced Data Mining Techniques*, Springer, 1st edition (February 1, 2008), page138,ISBN3-540-76916-1

[11] Chen Mosha,"Combining Dependency Parsing with Shallow Semantic Analysis for Chinese Opinion-Element Relation Identification", IEEE, 2010, pp.299-305.

[12] Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu,"Phrase Dependency Parsing for Opinion Mining", EMNLP '09 Proceedings of the Conference on Empirical Methods in Natural Language Processing, Volume 3,2009

[13] Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang,"Mining Product Reviews Based on Shallow Dependency Parsing", SIGIR '09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009

[14] Cost, R. S., Finin, T., Joshi, A., Peng, Y., Nicholas, C., Soboroff, I., Chen, H., Kagal, L., Perich, F., Zou, Y., and Tolia, S. 'ITTALKS: A Case Study in the Semantic Web and DAML+OIL.' IEEE Intelligent Systems 17(1):40-47, 2002.

[15] Davies, J., Weeks, R. and Krohn, U. 'QuizRDF: Search technology for the Semantic Web.' In WWW2002 Workshop on RDF and Semantic Web Applications, Hawaii, 2002.

[16] Deutsch, A.,Fernandez, M., Florescu, D., Levy, A. and Suciu, D. 'XML-QL: A query language for XML.' In Proceedings of the Eighth International World Wide Web Conference, 1999.

[17] Ding, L,. Lina Zhou, and Tim Finin, 'Trust Based Knowledge Outsourcing for Semantic Web Agents,' 2003 IEEE/WIC International Conference on Web Intelligence (WI 2003), October 2003, Halifax, Canada.

[18] Ding, L., Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Joel Sachs, Vishal Doshi, Pavan Reddivari, and Yun Peng, Swoogle: A Search and Metadata Engine for the Semantic Web, Thirteenth ACM Conference on Information and Knowledge Management (CIKM'04), Washington DC, November 2004.

[19] Marc Abrams, editor, World WideWeb:Beyond the basics, Printice Hall, 1998

[20] Mohd Wazih Ahmed, Dr. M. A. Ansari "A survey: Soft computing in Intelligent Information Retrieval Systems," International Conference on Computational Science and Its Applications, IEEE 2012

[21] S. Kalaivani, K. Duraiswamy, "Comparison of Question Answering Systems Based on Ontology and Semantic Web in Different Environment", Journal of Computer Science 8 (9): 1407-1413, 2012

[22] Hany M. Harb, Khaled M. Fouad, Nagdy M. Nagdy, "Semantic Retrieval Approach for Web Documents", (IJACSA) International Journal of Advanced Computer Science and Applications, 9, 2011

[23] Jianguo Jiang, Zhongxu Wang, Chunyan Liu, Zhiwen Tan, Xiaoze Chen, Min Li "The Technology of Intelligent Information Retrieval Based on the Semantic Web" 2nd International Conference on Signal Processing Systems (ICSPS), IEEE 2010

[24] Nicholas J. Belkin "Intelligent Information Retrieval: Whose Intelligence," Department of Information Studies, University of Tampere

[25] LIU Yong-Min, CHENG Shu "Artificial Intelligent Information Retrieval Using Assigning Context of Documents," International Conference on Networks Security, Wireless Communications and Trusted Computing, IEEE 2009

[26] Wenjie Li, Xiaohuan Zhang, Xiaofei Wei, "Semantic Web-Oriented Intelligent Information Retrieval System," International Conference on BioMedical Engineering and Informatics, IEEE 2008

[27] Yi Xiao, Ming Xiao, Fan Jhang "Intelligent Information Retrieval Model Based on Multi-Agents,", IEEE 2007

[28] Pan Ying, Wang Tianjiang, Jiang Xueling, "Building Intelligent Information Retrieval System Based on Ontology" The Eighth International Conference on Electronic Measurement and Instruments, IEEE 2007

[29] Tanveer J. Siddiqui, U. S. Tiwary "Integrating Notion of Agency and Semantic in Information Retrieval multi-agent model", Proceeding of the 2005 5th International Conference on Intelligent Systems Design and Applications(ISDA'05) , IEEE 2005

[30] https://en.wikipedia.org/wiki/Precision_and_recall