A Big Data Approach on Aspect based Summarization

Manjupriya R, Aarthi T
Department of Computer Science and Engineering
AVC College of Engineering,
Mayiladuthurai

Krishnakumari K
Associate Professor,
Department of Computer Science and Engineering,
AVC College of Engineering,
Mayiladuthurai

***Abstract*— In the present day scenario, selecting a good product is a cumbersome process. The reviews from the shopping sites may confuse the user while purchasing the product. It becomes hard for the customers to go through all the reviews, even when they read they may get into a baffling state. Some consumers may like to buy the best product based on its features and its extra comfort. Meanwhile, the size of the datasets for analysis process is huge which cannot be handled by traditional systems. In order to handle the large datasets, we are proposing a parallel approach using Hadoop cluster for extracting the feature and opinion. Then by using online sentiment dictionary and interaction information method, predict the sentiments followed by summarization using clustering. After classifying each opinion words, our summarization system generates an easily readable summary for that particular product based on aspects.**

Keywords: Sentiment analysis, summarization, aspect, hadoop

I. INTRODUCTION

Sentiment analysis is the area where sentences, words or documents are categorized as positive, negative or neutral based on the opinions expressed in the text. Many online companies wanted to increase their purchase rate by increasing the standards of positively reviewed products and tried to reduce the problems faced by the products which are negatively reviewed. People share their experiences immediately through social networks, blogs, short messages and many more means. The reviews given by each and every person is taken into consideration for further development of businesses. The contents like blogs, tweets, product reviews reflect the opinions of users depending upon the context. Most of the shopping sites are requesting for the reviews whenever they purchase the product. The interested people may share their experiences about the product purchased. People want to know the opinion of naïve users before purchasing any product. But it is difficult for any person to read all the reviews from a large number of reviews. People are also not satisfied with only a few good reviews.

Accordingly Opinion Mining or Sentiment Analysis followed by sentiment summarization plays a vital role in online shopping. It extracts the customers' opinion on each product and identifies the overall opinion along with a report telling experiences about the product. There are two ways of summarization of corpus which includes extractive summarization and abstractive summarization. Extractive Summarization is the strategy of concatenating extracts taken from a corpus into a summary while abstractive summarization involves paraphrasing the corpus using novel Sentences. In the summarization task, many words have the similar kind of orientation irrespective of their positions used. For example, the lexicons like “good”, “excellent”, “great”, “worst”, etc., have the same polarity as positive or negative based on the usage called as context-free words. But some other words are context based because they have different meaning in different places of usage. The sentiment polarity of context-based words can be identified by knowing the domain knowledge which is not an easy task to know the details about large number of domains. Here we propose an effective approach for identifying the contextually dependent words. Mining information from a massive dataset cannot be handled by traditional database systems. The Hadoop Map reduce framework is used to process large datasets using a cluster of commodity hardware.

II. RELATED WORK

An objective sentence explains the factual information about the product whereas a subjective sentence depicts the personal feelings, views or beliefs. The sentiment analysis can be performed at a document level, sentence level or at the aspect level. Aspect based sentiment analysis identifies sentiments on aspects of items. It can be either frequency-based where it searches for frequent nouns to identify the aspects or model based where it searches for model parameters. Sentiment analysis can be performed in different ways as i)Supervised approach ii) Semi-supervise approach or iii) Un-supervised approach. Supervised learning is the process of learning performed by mapping of labelled instances to output. But supervised approaches need large amount of training data which consumes more time for manual annotation. In the semi-supervised classification approaches, some amount of labelled data from the domain is considered for training the system. In the Un-Supervised approach, all the attributes are considered equal and independent where it groups the data based on some measure of resemblance. The semi-supervised and unsupervised approaches use the sentiment lexicon dictionaries like WordNet, SentiwordNet, SentiFul. Graph-based summarization framework called Opinosis which generates an abstractive summary of redundant opinions. Opinosis based summary contains short sentences and conveys the essential information. In 2015,

Moghaddam proposed a technique to extract actionable information from customer feedback. However, the author is looking only for defect or
Improvement. In the proposed work, we performed pre-processing, feature extraction using Hadoop, feature-opinion pair formation, opinion analysis and aspect based summarization. The summary can be presented in the form of opinion Table and opinion bar charts.

III PROPOSED WORK

Our summarization system handles context based opinion words in a scalable manner. We used Aspect based clustering method for summarization. We divide the entire tasks into five subtasks as

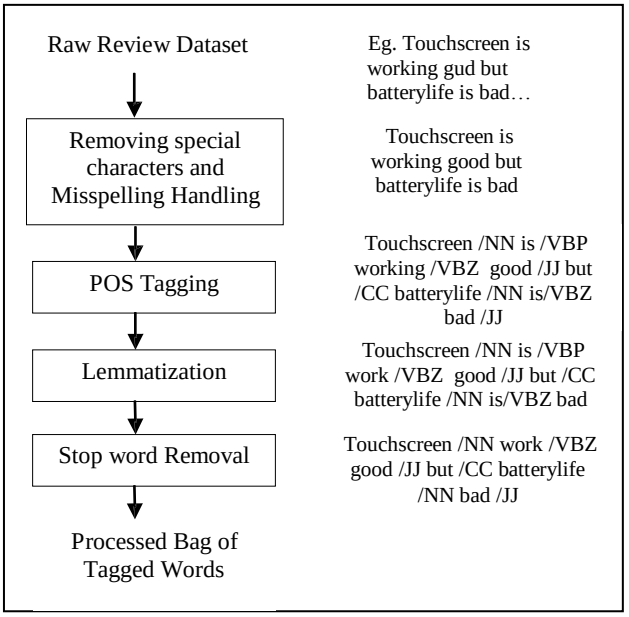
1. Pre-processing
2. Aspect Extraction in Hadoop Environment
3. Feature-Opinion pair formation in Hadoop Environment
4. Opinion Analysis
5. Aspect based Summarization

3.1 Pre-processing

The preprocessing can be performed in four main steps shown in Fig.1. After getting the review database, the first step is to handle the misspelled words and the special symbols used.
In the next step, we tag the words in the sentences using a Stanford POS tagger . In this step, each lexicon in the review document is tagged with their respective part-of-speech(POS). In the third step is lemmatization where the root word of the lexicon is identified. In the last step, we remove the unwanted tags which are not needed for further processing. We collect only the nouns, verbs, adjectives, adverbs and conjunctions.

3.2 Aspect Extraction in Hadoop Environment

After preprocessing we need to identify the important aspects. Based on the observations that most of the users give their review for important aspects and the reviews on important aspects influence the overall opinion of the product to a larger extent, aspect ranking can be done using aspect frequency. This can be done in Hadoop map Reduce environment to achieve scalability. Aspect identification is focused on extracting the set of aspects concerning the product from the reviews. It first identifies the nouns and noun phrases in the documents because aspects are usually in the form of noun and noun phrases. The occurrence frequencies of each noun and noun phrases are summed up and only the frequent ones are only considered for analysis purpose with the threshold count of ten. But at the same time the frequent nouns may not be the feature with respect to the product field (like defects, problems, comments, etc.,) The review „*my comments are always accurate with respect to iPhone reviews*", here the lexicon "comments" represents nouns but it is not a feature to be considered for the iPhone review summarization.



Fig(1) Steps in Pre-processing

3.3 Feature-opinion pair formation in Hadoop Environment

Based on the observations that most of the users give their review for important aspects and the reviews on important aspects influence the overall opinion of the product to a larger extent, aspect ranking can be done using aspect frequency. This can be done in Hadoop map Reduce environment to achieve scalability. Before going to map the opinion word with the nearest noun, we should handle the negation words because negation words may flip the polarity of the opinion word. The map and reduce processes of Mapreduce framework takes a key/value pair and outputs a key/value pair.Then construct the feature-opinion pair with the polarity obtained from SentiwordNet. For the classification of remaining Feature-opinion pair, we use the Interaction Information method for finding the co-occurrence between the aspect, opinion and the polarity. After classifying each opinion word, the system provides aspect based ranking for each feature in the product.

Feature	Modifier	Opinion
battery	Enough,very,not	Good,nice.
camera	Much,very,pretty	Worse,easy,great,good.
screen	Pretty,barely,very, fairly	Solid,visible,responsive, Receptive,good.
software	Rather,definitely,not	Easy,slow,flimsy,limited, Seamless.

Table 1:A Sample list of extracted Features, Opinions and Modifiers

Feature	Number of opinion words
phone	208
Apps	92
battery	70
screen	64
camera	37
interface	33
gaming	29
software	25
flash	14

Table 2. Top features in a sample of 145 Reviews

3.4 Opinion analysis

To identify the polarity of the context-based words, we use the linguistic rules of keywords like „and", „or", "neither", „nor", etc. The proposed approach uses equation (1) for calculating polarity of Feature-Opinion pair. The Contextual information (CI) for both positive and negative label is identified where the most relevant one is assigned as the label for the Feature-opinion pair.

$$CI(W,O,F) = \log_2 \frac{P(W,O)P(O,F)P(W,F)}{P(W)P(O)P(F)P(W,O,F)}$$

where W is the Opinion word,
O is the sentiment orientation label,
F is the feature associated with the opinion word.
Similarly P(W,O) represents total number of times W and O occurs together, P(O,F) represents total number of times O and F occurs together, P(W,F) represents total number of times W and F occurs together, P(W),P(O),P(F) represents total occurrence of W, O and F in the document.O is the sentiment orientation label, F is the feature associated with the opinion word.

3.5 Aspect based summarization

After classifying all feature-opinion pairs, we use aspect based clustering method for summarization. In which we used „2n" clusters for „n" feature. For each aspect, we create two clusters, one for positive and another for negative to separate positive and negative reviews. We extract the respective aspects based on the feature-opinion pair and put it in their respective cluster and count the number of reviews for positive and negative category.

We calculate the star rating of each aspect by dividing all the positive reviews by a total number of reviews represented by the feature. The final output will be the rating along with the top reviews of that product"s feature.

IV EXPERIMENTAL RESULTS

We performed the preprocessing in the Windows Intel Core Quad 2 CPU, 4 GB RAM with 1Gbps Network Speed against the Amazon datasets on phone domain. Then aspect extraction and feature opinion formation works are experimented in the Ubuntu 10.12 in the Windows Intel Core Quad 2 CPU, 4 GB RAM with 1Gbps Network Speed in a 2-Node Cluster.

4.1 Scalability

For a large number of reviews, our system provides good performance. To show the scalability of feature extraction, we change the number of reviews slowly in increased order and repeat the process a few number of times. The performance of the system increases gradually when increasing the number of reviews as shown in Fig.4. The file size ranges from 20 MB to
100 MB.

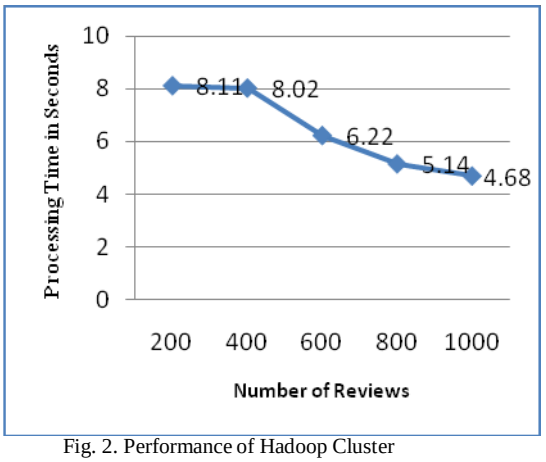
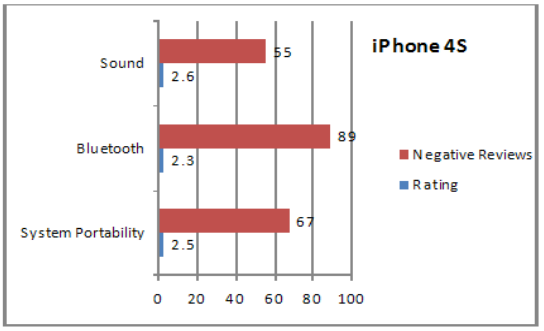


Fig. 2. Performance of Hadoop Cluster

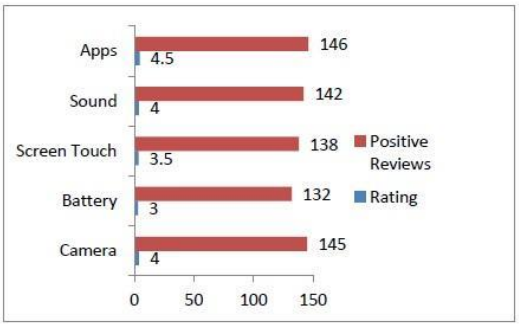
4.2 Comparison between different Aspects of iPhone – Positive Feedbacks

The features and opinions are extracted and then form the nearest pairs of each sentence of review documents. The top 4 features for positive reviews to appreciate and top 3 features for further development are listed using charts in Fig.3 and Fig 4.



iPhone 4S	System Portability	Bluetooth	Sound
Rating	2.5	2.3	2.6
Negative Reviews	67	89	55

Fig. 3. Negative Feedback Chart



iPhone 4S	Rating	Positive Reviews
Camera	4	145
Battery	3	132
Screen Touch	3.5	138
Sound	4	142
Apps	4.5	146

Fig. 4. Positive Feedback Chart

V SUMMARY OUTPUT

After clustering the reviews into two major groups as positive and negative, the consumers can get the summary information based on their respective aspects listed. A sample extractive summary is listed in Fig. 7 for the aspect *Apps*. When the user is interested to go through the complete review, they can then proceed with the links.

Sample summary for the aspect apps

it is super cool to get all the apps and the phone works great it fits in my purse nicely.....

big bright screen accurate multitouch unlimited flexibility through the app store.....

much faster than the original iphone great apps movies videos look fantastic.....

Fig. 5. Summary of the Project

V CONCLUSION

In this paper, we provide a simple and effective solution for scalable aspect based summarization. The co-occurrence framework (RMI) provides better results for identifying the context based words. The Hadoop cluster works well for feature-opinion extraction with large number of reviews. Our method provides an effective evaluative summary of those reviews without affecting the originality of the reviews. In future, an effective sentiment prediction method can be applied irrespective of their domains and then applying the summarization in the scalable environment with various presentation styles with the help of machine learning libraries like Mahout.

For future work, to improve the accuracy of our system by making feature extraction more accurate and including more opinion words, i.e., nouns, adverbs, etc. As including nouns and adverbs will lead to increase more non-opinion words, also try to create an effective mechanism that will select only opinionated words from nouns and adverbs.

VI REFERENCES

1. Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. InLREC (Vol. 10, pp. 2200-2204).
2. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008, April). Building a sentiment summarizer for local service reviews. In WWW Workshop on NLP in the Information Explosion Era (Vol. 14).
3. Briscoe, T., Carroll, J., & Watson, R. (2006, July). The second release of the RASP system. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 77-80). Association for Computational Linguistics.
4. Carenini, G., & Cheung, J. C. K. (2008, June). Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In Proceedings of the Fifth International Natural Language Generation Conference (pp. 33-41). Association for Computational Linguistics.
5. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
6. Ganesan, K., Zhai, C., & Han, J. (2010, August). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. InProceedings of the 23rd international conference on computational linguistics (pp. 340-348). Association for Computational Linguistics.
7. Ganesan, K., Zhai, C., & Viegas, E. (2012, April). Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. InProceedings of the 21st international conference on World Wide Web (pp. 869-878). ACM.
8. Gerani, S., Mehdad, Y., Carenini, G., Ng, R. T., & Nejat, B. (2014). Abstractive summarization of product reviews using discourse structure. In Proceedings of EMNLP.
9. Gindl, S., Weichselbraun, A., & Scharl, A. (2010). Cross-domain contextualisation of sentiment lexicons.
10. Hamid, F., & Tarau, P. (2015). Anti-Summaries: Enhancing Graph-Based Techniques for Summary Extraction with Sentiment Polarity. In Computational Linguistics and Intelligent Text Processing (pp. 375-389). Springer International Publishing.
11. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. InProceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.
12. Kansal, H., & Toshniwal, D. (2014). Aspect based Summarization of Context-based Opinion Words. Procedia Computer Science, 35, 166-175.
13. Kim, H. D., Ganesan, K., Sondhi, P., & Zhai, C. (2011). Comprehensive review of opinion summarization.
14. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer US.
15. Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013, October). Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier. InBig Data, 2013 IEEE International Conference on (pp. 99-104). IEEE.