ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

Features based opinion mining on online mobile products using data mining classification techniques

S. Arokia Mary[*1], P. Shanthi[2]

[1]*M.Phil.Scholar,* [2]*Associate Professor, Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu, India*
[1]arokiyamarry2016@gmail.com, [2]shanthimphil2016@gmail.com

Abstract

Objective: To extract sentiment or opinion words using repository of lexicons, to calculate overall polarity score for the product, to extract the aspects in the reviews, to devise polarity score for each aspects of the product and to develop a summary of the product aspects(targets) with its polarity score from the reviews.
Methods: There are several methods built up for sentiment analysis and opinion mining. In order to increase the recall, accuracy and precision openNLP parser with naïve bayes classifier is proposed. The opinion lexicons are used to produce summary about the reviews.
Findings: Opinion mining is a challenging Natural Language Processing or text mining problem. The reason behind this is we can't exactly decide what user says about particular product. Because each one's writing style would be different. All the reviews expressed in the websites cannot be processed directly. The reviews must be preprocessed in order to eliminate unnecessary characters. Many techniques were proposed for opinion mining but it lacks in accuracy, precision and recall.
Applications/improvements: To enhance the accuracy of opinion mining openNLP parser with naïve bayes classifier is proposed.

Keywords: openNLP parser, polarity classification, sentiment analysis, opinion mining.

1. Introduction

Opinion mining [1] is the field of study which discusses about people's emotions, sentiments and evaluations about anything. Basic technologies and tools used by Opinion Mining are web search engine, information extraction, data mining and information retrieval. Information retrieval is the process of obtaining information relevant to the needed information. Information extraction is the act of extracting structured information from the web. People express their views and opinions on the web. They can now post their reviews about product or opinion about anything on the discussion forums, blogs, social websites, or at merchant sites. Opinions expressed or posted by the people called as user generated contents.

In an existing system [2], sentiments were analyzed and opinions were mined using NLP. This system used Twitter Auth for streaming tweets. Then the data were preprocessed that converts the unstructured data into structured data. Based on some criterion apply weights to tweets and also rate of tweets on film hash tag. This utilized Non Language processing for preprocessing. This has some issues like huge computing power and precision problem. This is overcome by using our proposed technique openNLP parser with naïve bayes classifier (data mining techniques). The experimental results prove that the proposed technique shows high recall, accuracy and precision than the existing technique.

In [3] proposed an approach to handle and classify objective and subjective sentences in the unstructured data for sentiment analysis [4]. It analysed various different approaches like machine learning techniques, text mining techniques, natural language processing, techniques of information theory and coding and semantic and hybrid approach for sentiment analysis of unstructured data in reviews. The proposed work considered objective sentences where it classified the documents as opinionated and non opinionated. Further opinionated sentences were classified as objective or subjective sentences. Then the subjective and objective sentences were classified based on their polarity. However the threshold value in classification may the performance of sentiment analysis.

In [5] proposed an approach which combined both neural network and naïve bayes classifier for sentiment classification. From the analysis of data mining technique naïve bayes it is found that it lacks accuracy for complex real world applications. But the neural network effectively manages the correlation between the input variables. Thus, the naïve bayes and neural network were combined together to improve the performance of sentiment classification. However it failed to address the problem of scarcity of opinion annotated data in a language.

In [6] analyze the sentiments in review documents to predict movie's box office success. The success of a movie was predicted by using simple metrics. In order to predict the success of a movie analyzed sentiments in Tweets during the movie release. Initially the data was collected from existing twitter data set then the collected data are pre-processed using filters to convert the input data into suitable format to sentiment analyzer tool. Finally Ling pipe sentiment analyzer is used to classify the sentiments. However, the temporal analysis is not considered in this sentiment analysis.

In [7] presented a linear regression model to predict the future outcomes of movies. In this paper analyzed the sentiment presented in the reviews in Tweets. The regression model has strong correlation between the input variables or amount of attention of a given topic and its ranking in the future. However, the prediction accuracy is not attained to expected level.

In [8] proposed solution for rated aspect summary problem. The proposed solution is generated by describing the rated summary problem and provides solution which is decomposed into three steps. The first step wasextracting major aspects from the reviews. Then the second step was predicting rating for each aspect from the overall rating. Final step was extracting representative phrases. However, this solution is not incorporate with how to compare entities in the reviews more effectively based on the rated aspects.

In [9] introduced and analyzed sentiment analysis and opinion mining. The various machine learning techniques like maximum entropy, Support Vector Machine and naïve bayes were utilized for opinion mining. The sentiment analysis based on bag of words focused on context involved in reviews. Finally in this paper concluded that the support vector machine with the kernel methods were more efficient for sentiment data classification.

In [10] proposed an approach for automatic sentiment classification. This proposed approach counts the number of positive and negative words in reviews to find the sentiment orientation and features are extracted using SentiWordNet lexical resources. Then the extracted features are given as input to classifiers in order to classify the sentiment words in the reviews. From the analysis it was found that this SentiWordNet is similar to manual lexicons. However, it has limitation like SentiWordNet's reliance on glosses.

In [11] proposed an algorithm for sentiment analysis which replaced the traditional stochastic gradient descent algorithm in tree leaves. In this paper, the proposed work has two phases: pre-processing and online processing. Initially pre-process the data and apply genetic algorithm based on Bayesian approach for dimensionality reduction; then finally apply Hoeffding's Stochastic Gradient Descent Tree for sentiment classification. However, weighted threshold may affect the performance of sentiment classification.

2. Materials and methods

The review dataset contains reviews posted by the customers on the blogs, forums and commercial websites. These reviews are pre processed for removing unnecessary symbols or text present in the review dataset. For example the reviews may contain product id, review id like # symbols etc. During preprocessing step these are cleaned in order to obtain good quality of results when extracting and opinions and aspects.

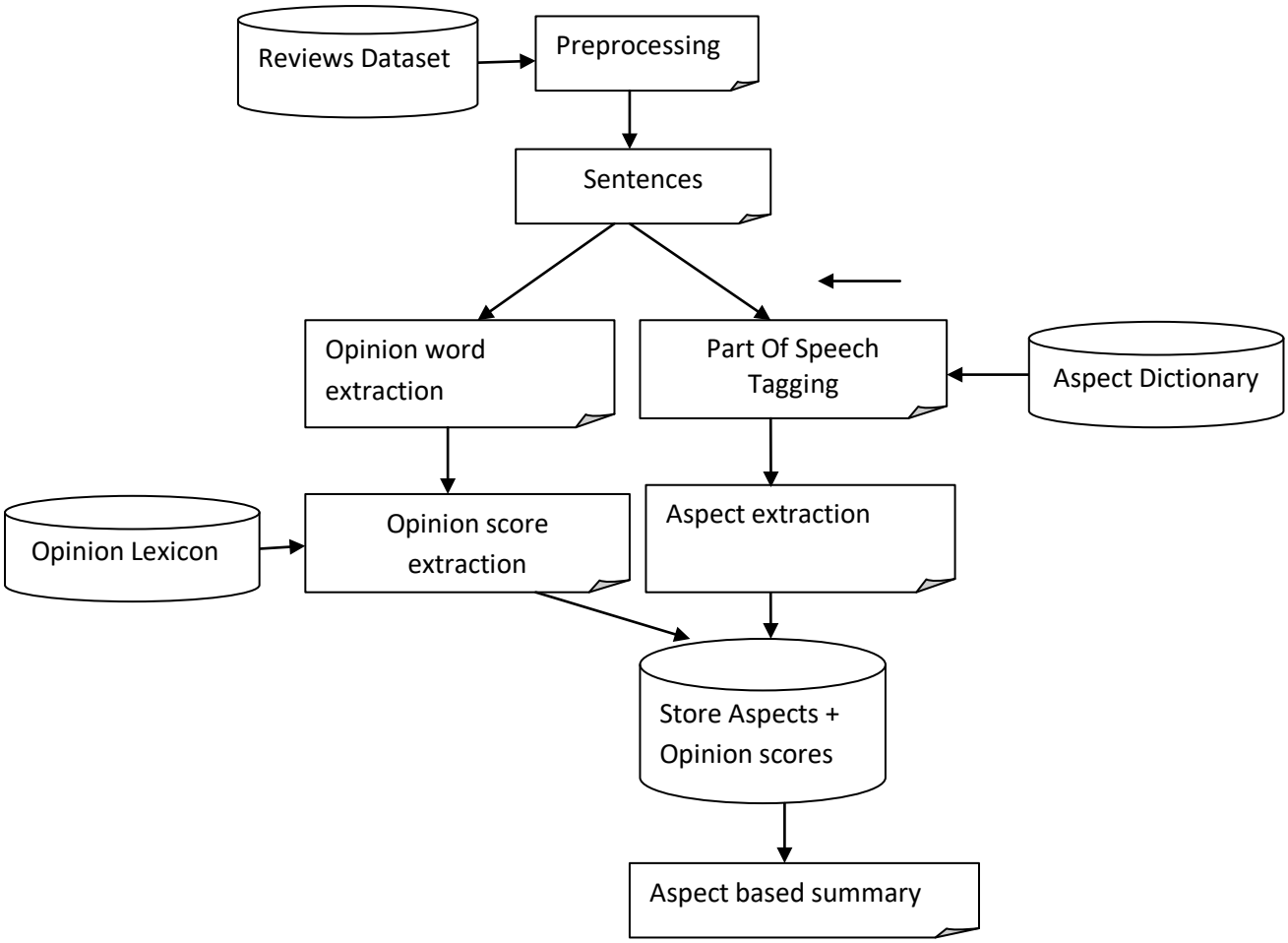
After preprocessing we get only the sentences extracted from dataset. Using Hu and Liu's opinion lexicon we extract the opinion words with its polarity score for each sentences. And then we use openNLP Part of Speech tagging technique to extract aspects from the sentences. Once aspect present in the sentences extracted the sentence score is considered for the aspect. Finally these values are stored and used for producing aspect based summary. The technique used by our proposed approach is given in detail in the following sections. Figure 1shows the overall architecture of proposed system.

2.1. Part of speech tagging for aspect extraction

By analyzing the natural language we can determine the likes and dislikes of users of a particular product through the POS tagging. POS tagging is the partition of speech tagging from natural language processing. The opinion of a product is determined by extracting or finding the features of product using POS tagging.

"The mobile's sound quality is pretty good."
Here the product user is talking not only about the product mobile but mentions the product feature "*sound quality*".

Figure 1. Proposed Work for Aspect and Opinion Extraction



The POS is a category utilized in linguistics. The linguistics is described by a morphological or syntactic behaviour of a word. As we know in the English language grammar the part of speech are classified as pronoun, verb, preposition, noun, adverb, interjection, adjective and conjunction. Additionally more categories of part of speech are mentioned latter like articles and even more based on variations of the aforementioned ones.

The main reason of considering POS tagging in information extraction is the each category of POS plays a pivotal role within the sentence. Nouns in sentence given names to objects, entities or beings. An adjective in a sentence describes or qualifies nouns. Similar to adjective adverbs defines the nouns in a sentence but in different conditions or situation it rarely modifies a noun. The following are some examples of part-of-speech which are taken from opinions on the internet:

i) My Samsung mobile is bad
ii) I have a good Samsung mobile.
iii) I always have problems with the sound system.
iv) This mobile comes with a recharger that works everywhere.
v) The sound system never works.

From the above opinions i and ii describes the noun Samsung mobile as bad and good respectively. In iii, iv and v the words in the sentence are always, Ini and ii, both good and bad qualifies the noun mobile. In iii, iv and v the words are always, everywhere and never qualify the noun but different from adjectives they modify the verb. The likelihood of a word to belong to a specific group is identified by utilizing POS tagging in various techniques of data mining.

The part-of-speech has morphological and syntactical behavior by tokenizing a sentence with their part-of-speech and then developing hooks for various algorithms to takeout certain words in a sentence. As significance, algorithms can be more efficient and less error prone. Finally, an algorithm infers semantics in a sentence to develop intelligent associations.

Our main aim is to find out the features that explicitly appear as noun phrases or nouns in the reviews. The noun or noun phrases in a review was determined by using Part-of-speech. Hence in the proposed work, to find out the adverb, noun, adjective, verb in a review an effective parser called openNLP parser is used. The openNLP parser used to display sentence with the POS tags.

2.1.1. Extracting aspects from POS

Our proposed approach extracts the part of speech of each sentence using openNLP parser. We extract Noun/Noun phrases from the POS extracted before from review sentences. Also we have developed am manual aspect list for our product. Thus we match the aspect list with the Noun (NN) list extracted. If the aspect is found in a review sentence then the aspect related opinion words exists is extracted using opinion lexicon.

2.2. Building Aspect list for Feature Extraction

And we also generate manual list of aspects for the product. An example of mobile aspect list is shown in figure 2.

Figure 2. Aspect list for mobile

Mobile Aspects:

Battery, Screen, Camera, Radio, keypad, Applications, Speed, Bluetooth, Display, Speaker, etc.,

Aspect extraction algorithm

Aspect extraction algorithm extracts noun from Part Of Speech extracted from review sentences using openNLP parser. Figure 3 shows the algorithm for aspect extraction from review sentences.

Figure 3. Aspect extraction algorithm

```
AspectExtraction (Sentence S)
P :=POS-Tagging(S)
    Noun:=nouns(p)
    If (Noun is not empty)
    A:=A U Noun
return A
```

2.3. Naive Bayes approach for Polarity classification

In this section, the process of finding polarity of each sentence is described, which decides whether the review is positive, negative or neutral. For score extraction we utilized an efficient supervised learning technique called naïve bayes approach. The naïve bayes approach find out the positive and negative word with the help of AFINN word list. The process is used extract the polarity words from the sentence using opinion lexicon which includes the negative and positive words from the dictionaries.

i. Split the reviews into sentences and assign unique ID to each review. Clear noise from text and apply POS by removing all hash tags (#topic), targets (@username), and special Words. POS tagger is used to tag adverbs, verbs and adjectives.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

ii. Check each word of a sentence in the Opinion lexicons, Aspect dictionary. If it present then the sentence is labeled as opinion or subjective sentence.

iii. If the extracted Noun/Noun Phrases exist in Aspect lexicon then retrieve its polarity count from the polarity list we have already extracted.

iv. And finally we summarize the aspect and its corresponding polarity extracted from stored

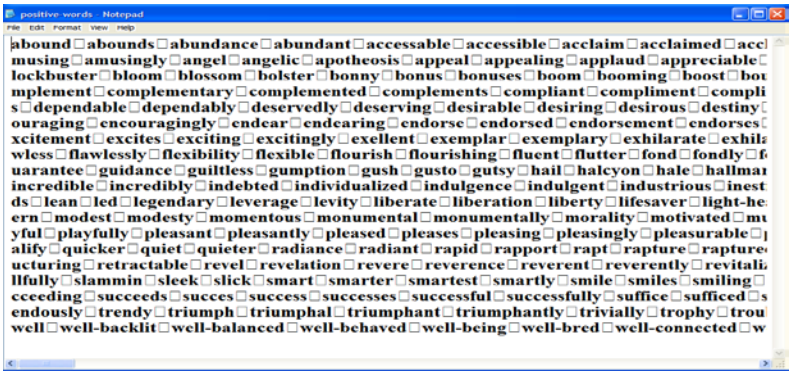
2.3.1. Opinion lexicon for polarity score extraction

The opinion lexicons contain opinion words expressing positive or negative sentiment. The words that depict a desirable condition such as good, nice, fantastic have positive polarity, while words expressing undesirable state have negative polarity (e.g., bad, awful, dreadful). The opinion lexicon also stores polarity scores of such positive and negative words. This list was prepared by Hu and Liu in the year 2004.

Positive Opinion Words

Here is the positive opinion lexicon shown in figure 4

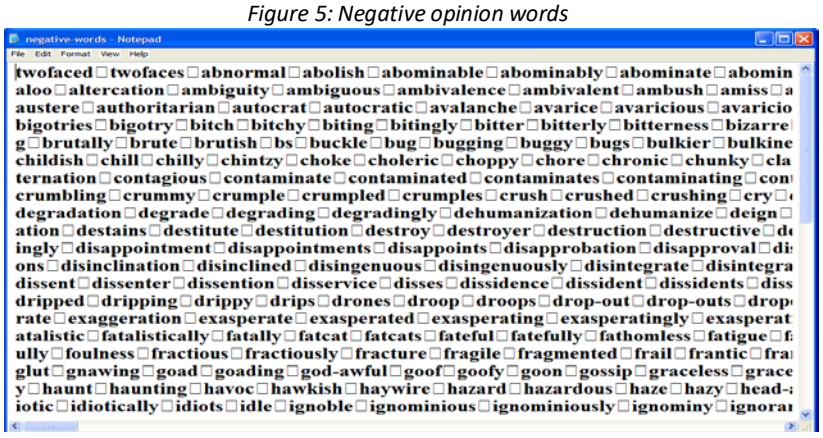
Figure 4. Positive opinion words



Negative opinion words

The Figure 5shows that a text file contains only negative word list.

Figure 5: Negative opinion words



2.4. Producing Review summary based on extracted opinion and Polarity Score

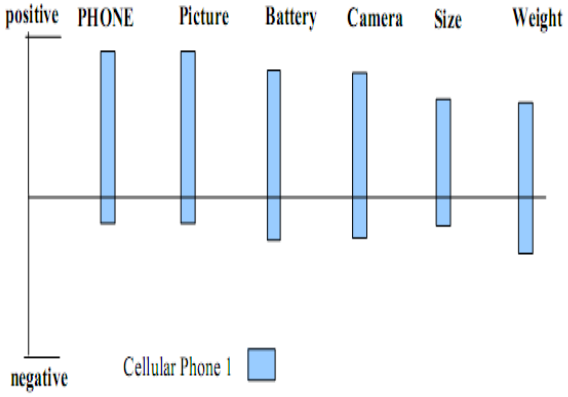
Opinion summarization is the final step of aspect based opinion mining. Opinions collected from multiple customers are analyzed and aspects and its polarity expressed by the users are extracted. Finally using the results we have got in the previous process we produce the aspect based summary. This summary includes two main characteristics. First, the opinion aspects (or features) and its polarity expressed by the user. Second, the quantitative summary of the number of people who hold positive or negative opinions about the product aspects(or features). The resulting opinion summary is a form of structured summary produced from the extracted information. Most opinion summarization methods produce a short text summary which includes the positive and negative count for the specific aspect. Here is a sample opinion based summary produced for the product mobile is shown in figure 6 and figure 7.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

Figure 6. An aspect-based opinion summary.

Mobile phone:
Aspect: GENERAL
Positive: 125<Individual review sentences>
Negative: 34<Individual review sentences>
Aspect: Speaker
Positive: 75<Individual review sentences>
Negative: 25<Individual review sentences>
Aspect: Battery life
Positive: 50 <Individual review sentences>

Figure 7. An example of aspect based summary



3. Results and discussions

The results of the existing and the proposed opinion mining are performed and in order to prove the effectiveness of the proposed opinion mining with different measures like accuracy, precision and recall is compared with existing technique. For this experimental purpose twitter datasets about mobile products are used.

3.1. Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.
Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

From the figure 8, it is proved that the proposed technique has high accuracy than the existing technique. In the graph x axis represents methods and y axis represents accuracy. The proposed technique has accuracy of 88.23 and the existing technique has accuracy of 76.4.Table 1 shows the accuracy comparison of proposed and existing technique.

Figure 8. Accuracy comparison

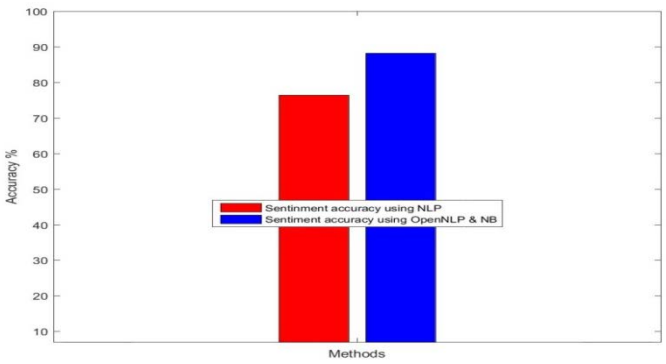


Table 1. Accuracy comparison

	Sentiment accuracy using NLP	**Sentiment accuracy using NLP & Naive Bayes classification**
Accuracy	76.4%	88.23%

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

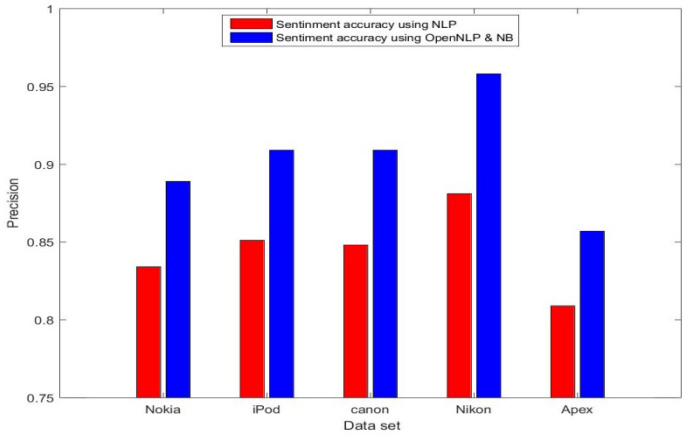
3.2. Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$Precision = \frac{Truepositive}{(Truepositive + Falsepositive)}$$

From the figure 9, it is proved that the proposed technique has high precision than the existing technique. In the graph x axis represents datasetsand y axis represents precision. The proposed technique has precision of 0.889 for Nokia dataset, 0.909 for iPod dataset, 0.909 for canon dataset, 0.958 for Nikon dataset and 0.857 for Apex dataset but the existing technique has precision of 0.834 for Nokia dataset, 0.851 for iPod dataset, 0.848 for canon dataset, 0.881 for Nikon dataset and 0.809 for Apex dataset.

Figure 9. precision comparison

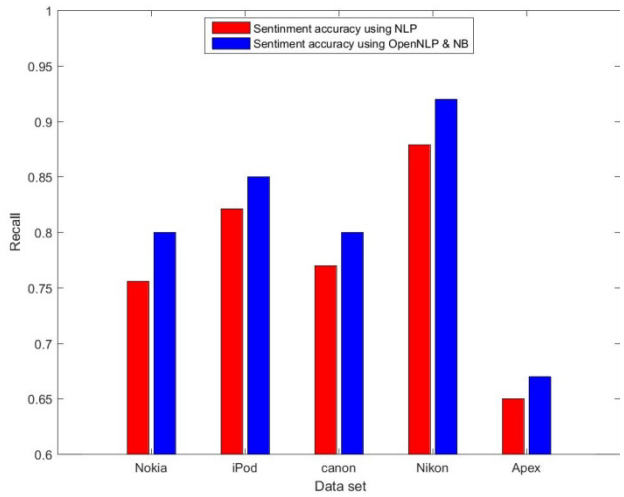


3.3. Recall

The Recall value is evaluated according to the classification of data at true positive prediction, false negative.

$$Recall = \frac{Truepositive}{(Truepositive + Falsenegative)}$$

Figure 10. Precision Comparison



From the figure 10, it is proved that the proposed technique has high recall than the existing technique. In the graph x axis represents datasets and y axis represents recall. The proposed technique has precision of 0.8r Nokia dataset, 0.85 for iPod dataset, 0.8 for canon dataset, 0.92 for Nikon dataset and 0.67 for Apex dataset but the

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

existing technique has precision of 0.756 for Nokia dataset, 0.821 for iPod dataset, 0.77 for canon dataset, 0.879 for Nikon dataset and 0.65 for Apex dataset.

Table 2. Opinion summary for web products

S. No	Product name	Aspects	Opinions
1	Nokia	Battery	good
2	Nokia	Camera	Compact
3	Nokia	Price	High
4	Nokia	Picture	Bright
5	Nokia	Speed	High
6	Nokia	Speaker	Good

Table 2 discusses the results we obtained based on particular aspects of a product as a summary. This summary is analyzed from reviews sentences and the maximum positive or negative is taken to produce a summary for each aspect.

Results can be evaluated using standard Information Retrieval (IR) metrics Precision and Recall respectively. Table 3 contains the results for different products.

The Table 4 is the summary we got for mobile product aspects. It gives the details of how many people have positive opinions and how many people have negative opinion about different aspects of the product.

Table 3. Classification of opinion aspects extraction for web (online) products

S. No	Product name	Total no. of aspects	Aspects extracted (POStags)	Correct aspects	Precision (correct aspects/extracted aspects)	Recall (correct aspects/Total aspects)
1	Nokia	40	36	32	0.889	0.8
2	iPod	35	33	30	0.909	0.85
3	canon	25	22	20	0.909	0.8
4	Nikon	25	24	23	0.958	0.92
5	Apex	18	14	12	0.857	0.67

Table 4. Positive and negative opinion for product aspects

S. No	Product Name	Aspect	Positive	Negative
1	Nokia	Battery	30	6
2	Nokia	Speed	20	5
3	Nokia	Camera	29	8
4	Nokia	Speaker	12	3

4. Conclusion

An aspect based opinion mining technique allows us to analyze opinions about product aspects(or features) that are related with the product such as product attributes. This thesis extracts aspects and the related customer opinions (or sentiments) on the online (web) product domain. It results in an approach which discovers customer preferences about product aspects. The proposed system allows summarizing the information obtained in order to provide a clear cut view of a product to the customers. Since the reviews collected were from the previous users, this summary helps the users to take a decision before buying a product online. Also this product summary helps the manufacturers to improve their product quality.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

5. References

1. Bing Liu. Sentiment analysis and opinion mining. synthesis lectures on human language technologies. *Morgan & Claypool Publishers*. 2012.
2. S. A. Mulay, S. J. Joshi, M. R. Shaha, H. V. Vibhute, M. P. Panaskar. Sentiment analysis and opinion mining with social networking for predicting box office collection of movie. *International Journal of Emerging Research in Management and Technology*. 2016; 5(1), 74-79.
3. J. S. Modha, G. S. Pandi, S. J. Modha. Automatic sentiment analysis for unstructured data. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2013; 3(12), 91-97.
4. T. SivaKumar, Amitha Joseph. A survey on different approaches for sentiment analysis of people. *Indian Journal of Innovations and Developments*. 2016; 5(6), 1-4.
5. L. L. Dhande, G. K. Patnaik. Analyzing sentiment of movie review data using Naive Bayes neural classifier. *International Journal of Emerging Trends and Technology in Computer Science*. 2014; 3(3), 313-320.
6. V. Jain, Prediction of movie success using sentiment analysis of tweets. *The International Journal of Soft Computing and Software Engineering*. 2013; 3(3), 308-313.
7. S. Asur, B. A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).* 2010; 1, 492-499.
8. Y. Lu, C. Zhai, N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wideweb, ACM*. 2009; 131-140.
9. J. Khairnar, M. Kinikar. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*. 2013; 3(6), 1-6.
10. B. Ohana, B. Tierney. Sentiment classification of reviews using SentiWordNet. In *9th. IT & T Conference.* Ireland. 2009; 13.
11. S. S. Minab, M. Jalali, M. H. Moattar. A new sentiment classification method based on hybrid classification in twitter. *2015 International Congress on Technology, Communication and Knowledge (ICTCK),IEEE*. 2015; 295-298.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

S. Arokia Mary, P. Shanthi. Features based opinion mining on online mobile products using data mining classification techniques. *Indian Journal of Innovations and Developments.* 2016; 5 (9), September.