Product Attribute Sentiment Analysis

Raju B N
Teaching Faculty,
Department of CSE,
KSWU, Vijayapura.

Rajahmad Jumnal
Teaching Faculty,
Department of CSE,
KSWU, Vijayapura.

Santosh Pawar
Teaching Faculty,n
Department of CSE,
KSWU, Vijayapura.

ABSTRACT
In the digitized world today internet is tone of the main sources of the information. There are several e-commerce websites where people/customers discuss different aspects/issues of the product. All such website provides a platform for the consumers to discuss and provide their opinion about the product, its features and their services. These opinions and reviews of the consumers provide very rich information both for other users as well as firms. But the issue with this information is that the information is mostly unorganized and therefore it is difficult to create a knowledge base out of it. What this paper propose is a product facet ranking framework which automatically determines the important facets of the product from the online comments/reviews, aiming to improve the usability of the consumer reviews. The facets are identified by the following observations; 1) The important facets of a product are usually given by several consumers in the review. 2) The review/opinion of the consumer on important facet of the product influences the overall opinion or view of the consumer on the product. But identifying the most important facets will increase the usefulness of the innumerable reviews/opinions and is useful to both users and the firms itself. It is practically difficult for the people to manually identify the important facets of the products from the consumer reviews/opinions. Consumer can easily make purchasing decision by focusing more on the important features, while the firms can concentrate on improving the quality of the features or facets of the product.

Keywords
Keywords are your own designated keywords which can be used for easy location of the manuscript using any search engines.

1. INTRODUCTION
In today's world internet and being online plays a very important role in day-today life. It has become a very popular communication tool among Internet users. Many millions of messages, views are appearing daily in popular e-commerce websites that provide product and services for consumers. Internet provides simple and easy way for the people to do multiple tasks of their everyday life like navigating through multiple websites, online purchasing of products, online transactions etc. Due to recent increase in the access of internet availability and it's penetration to even the remotest corner of the country has resulted in more number of products being sold on the internet. Several thousands to millions of products from different dealers are offered online. For example, Flipkart has several thousands of products from several dealers; Amazon.com registers more than thirty to thirty five million products. Thus it affects several more number of people trying to buy the products online. From the last few years there is a fast growth and the emergence of the e-commerce technology has inspired the costumers to buy the products online and express their opinion and write reviews about the product features and services online. To earn the trust of the customer the firms allow the customer to write the reviews/opinions about the features of the product and their online experience about the website and the firm. This has resulted in large number of people writing their views online and the product feature reviews has grown rapidly. For ex. "Issue in Yu Yureka Plus phone: Heating issue and touch is not working properly." of product Yu Yureka Plus mobile phone.

These reviews/opinions by the users are very important for the firms and they can help in increasing the sale of their product and can also be used for the brand building for the firm. The reviews/opinions are also helpful for other users to make intelligent decisions about buying a product online and also help the merchants/firms in knowing what are the positive and/or negative features of their product. A sentiment is the opinion expressed by the customer or a user. Sentiments represents the likes, dislike or aversion and some time the neutral view about the product or it's one or more features. This sentiment classification can be done at different levels such as Document level, Sentence level, and Attribute level. Document level sentiment analysis is useful when reviewing blogs. Sentence level sentiment analysis is useful while extracting it from reviews/opinions written by the user on websites. Consumers usually find such quality reviews very useful while buying over the internet. Many companies use the online reviews/opinion as feedback of the customers about their products and it's features. They can also use the reviews for marketing and CRM (Customer Relationship Management). These reviews can be used by the firm in improving the quality of the product and its features.

In this paper it is proposed that, first recognize the features or the attributes of the product then classify the sentiment on the attribute and then provide a rating to the product..

2. LITERATURE SURVEY
Here the paper review the existing works relating to the framework that is being proposed. This paper will start with the works that has been done on the attribute identification from the user review. Existing methods for attribute identification based on the lexicon-based approach and supervised learning approaches.

The lexicon-based approach uses dictionaries of words with their explanations of semantic orientation and includes the intensification and negation. The supervised learning is a machine learning approach that uses the training data consisting of a set of training examples. There are many learning-based classification models such as Naive Bayes, ME model, SVM etc. Supervised learning cannot perform efficiently without sufficient training samples. However labelling the training data is not easy and is also very time consuming. Here the pros and cons have to be explicitly categorized as positive and negative reviews/opinions on the attributes. Jin and Ho [1] learned a lexicalized HMM model to extract aspects and opinion expressions, while Li et al. [2]

integrated two CRF variations, i.e., Skip-CRF and Tree-CRF. The most remarkable unsupervised approach was proposed by Hu and Liu [3]. They assumed that the product attributes are nouns phrases and noun. The approach extracts nouns phrases and noun as potential candidate attributes. The occurrence frequencies of the nouns phrases and noun are computed, and only the repetitive ones are kept as attributes. The next step is attribute sentiment classification, which controls the orientation of sentiment expressed on each attribute. The lexicon-based methods are usually unsupervised. They rely on a sentiment lexicon containing a list of negative and positive sentiment words. To produce a high quality lexicon, the bootstrapping strategy is usually employed. For example, Hu and Liu [3] started with a set of adjective seed words for each opinion class (i.e., positive or negative). They utilized antonym / synonym relations defined in WordNet to bootstrap the seed word set, and finally obtained a sentiment lexicon.

Document-level sentiment classification aims to classify an opinion document as expressing a positive or negative opinion [7]. Present works use unsupervised, supervised or semi-supervised learning techniques to build document level sentiment classifiers. Those methods that are unsupervised usually depends on a sentiment lexicon containing a stock of positive and negative sentiment words. It controls the overall opinion of a review document based on the number of negative and positive words in the review. Supervised method applies existing supervised learning models, such as SVM and Maximum entropy (ME) etc. [4], while semi supervised approach exploits abundant unlabeled reviews together with labelled reviews to improve classification performance. The two widely used methods are the sentence ranking and graph-based methods [5]. In these works, a scoring function was first defined to compute the informativeness of each sentence. Sentence ranking method [2] ranked the sentences according to their informativeness scores and then selected the top ranked sentences to form a summary. Graph-based method [6] represented the sentences in a graph, where each node corresponds to a sentence and each edge characterizes the relation among two sentences. A arbitrary walk was then performed over the graph to find the most informative sentences, which were in turn used to construct a summary.

Example: Cellular phone 1: MOBILE

+ve: 125	<individual review sentences>
-ve: 7	<individual review sentences>

Attribute: voice quality

+ve: 120	<individual review sentences>
-ve: 8	<individual review sentences>

Attribute: size

+ve: 80	<individual review sentences>
-ve: 12	<individual review sentences>

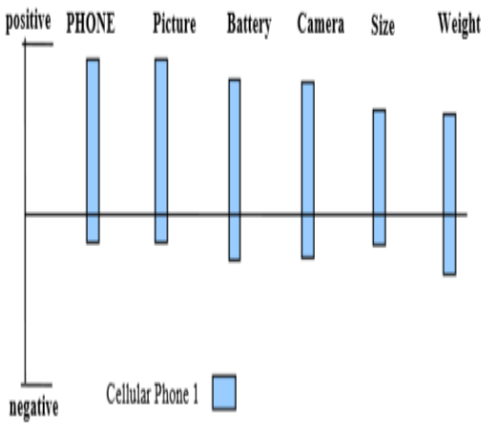

Fig 2.1 (A): Visualization of an attribute based summary of opinions on a cellular phone

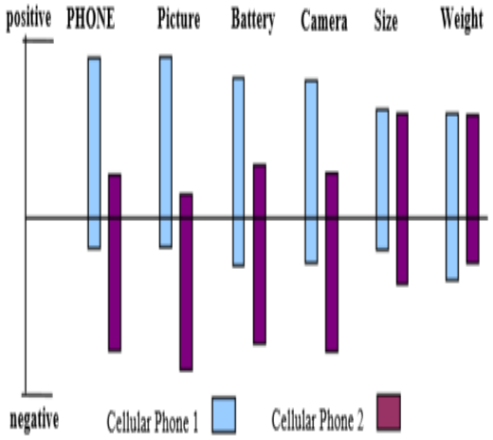

Fig 2.1(B): Visual opinion comparison of two cellular phones

3. PROPOSED WORK

Below mentioned 4 steps are used in the proposed work.

1) Reviews/opinion extraction and Pre-processing.

2) Identification of the product attributes

3) Classifying the positive and negative reviews/opinions of the product by sentiment classifier.

4) The prospect ranking algorithm used to rank the products.

Introducing the required tools,

1. SentiWordNet:
This paper propose to use SentiWordNet 3.0 for having the value of the user's opinions or sentiments. Basically it is using the negative and positive values. For the 0's of both (+ve and -ve), this paper considers the neutral opinion.

2. Twitter website:
From twitter website (www.twitter.com) one has to extract the tweets to analyze them. Ex: In twitter one can search for any topic or product. Different people have different thoughts of their products and they share their +ve and -ve opinions. These tweets are very useful for the analysis.

3. Python (Eclipse 3.7.1):
For developing applications in Python, Eclipse is the well suited software development environment. It comprises with IDE and various plug-ins.

4. Databases:
Authors are using the databases in storing tweets after extracting and also the SentiWordNet words with the values. Databases are in the form of MS Excel files (such as xls or csv).

3.1 Reviews/opinion extraction and Pre-processing.

The first step of the product attribute ranking structure is data pre-processing which is very important. Analogizing to traditional text document the reviews are generally less traditional and written in specific manner. If the sentiment analysis is applied on an unstructured review/opinion generally achieve poor performance in majority of the case. Therefore the pre-processing techniques on reviews/opinions are very important to obtain the result of sentiment analysis to be at acceptable level

3.1.1 Searching Tweets

Authors are using the micro-blogging website twitter.com to have the tweets from the user. To get the tweets authors are using Twitter API. The users can post their view or opinion regarding the attributes of the product depending on their ideas or their liking about any product or service. Authors propose to use the Twitter API to for retrieving tweets from the Twitter website and at a time authors are extracting tweets.

Once one get the attributes one can classify them in two forms,

a. Subjective and

b. Objective

Subjective means it is portrays of the mixture of +ve and -ve values or sets and on the other hand, the other hand objective means representation of neutral values (0) or sets. Authors are using POS tagging where authors used the parts-of-speech[11] for tagging the tokens. Subjective texts tend to use base form of verbs (VB) and also simple past tense (VBD) instead of past principle (VBN). Adverb (RB) is mostly used in subjective texts to give an emotional colour to a verb.

3.1.2 Saving the extracted Tweets

After extracting the tweets authors have saved the tweets in a text file which is in JSON form. It is then converted to .csv form.

3.1.3 Tagging the Tweets

For sentiment analysis authors first tag the tweets according to the POS such as noun, verb, adjective etc. As preposition and conjunctions are too common to be used among the sentiments, one can easily remove them after tagging as well as proper nouns which usually don't have an effective content. After removing the un-affecting content, one usually have 4 POS: adjective, noun, verb and adverb which are known as opinion words.

Part-of-speech tagger is an application that helps to read text in some language and assign POS to individual terms, words or tokens. The tagger that has been used here is written by The Stanford Natural Language processing. Authors propose to use the English tagger model. Some of the taggings are as follows.

VBN -> Verb, past participle

VBZ -> Verb, 3rd person singular present

RBS -> Adverb, superlative

WPS -> Possessive wh- pronoun

POS -> Possessive Ending

NNS -> Noun Plural

IN -> Preposition

JJR -> Adjective, comparative

MD -> modal

TO -> to

NN -> Noun, singular or mass

JJS -> Adjective, superlative

RP -> Participle

VB -> Verb base form RBR

RBR -> Adverb comparative

JJ -> Adjective

VBG -> Verb, gerund

DT -> Determiner

VBD -> Verb, past tense

WDT -> Wh- determiner

WP -> Wh- pronoun

CC -> coordinating Conjunction

EX -> Extential there

PDT -> Predeterminer

UH -> Injection

VPB -> Verb, non 3rd person

WRB -> Wh- adverb

RB -> Adverb

Following are the methods of pre-processing,

1) Stemming: Here one will remove the prefix from each word such as ing, tion, sion etc Eg 1. Running will become Run after stemming. Eg 2. "arguing", "argue" and " argued" truncated to the stem "argu"

2) Tokenization: It is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. This is related to sentiment analysis as referring to Senti-WordNet works with those tokens. Here one have to remove all the spaces, stop-words like a, an, the, are etc.

3.2 Identification of the product attributes

Attribute identification of the product is the next important step in which one can recognize the attributes from numerous customer reviews/opinions. The reviews are available on several different e-commerce web-sites. The issue with these reviews is that they are collected in different formats on different web-sites. Customer reviews/opinions contain positive, negative and also neutral reviews/opinion. Some web-site does overall rating of the product and some other do in the form of paragraph. One can get precise aspect by

extracting persistent noun from the positive and negative reviews [8]. Hu and Liu proposed most remarkable approach for attribute identification. In this approach it identifies the noun phrases and noun in the review/opinion. The occurrence frequency of noun phrase and noun are counted then only the consistent noun terms are kept as the aspect [3]. Some aspect may contain synonym term such as "earphone" and "headphone". In such situation One can perform synonym clustering to obtain unique aspect. These synonyms are obtained from synonym dictionary web-sites [7].

3.3 Classifying the positive and negative reviews/opinions of the product by the sentiment classifier.

Sentiment analysis is a type of NLP (Natural Language Processing) that is used for tracking the feeling or/and the polarity of the user /customer about the product and it's attribute. Sentiment classification [7] is used to classify the given word or sentence to one or more predefined sentiment categories such as Positive, Negative. There is various classification techniques are available. There are two types of learning supervised learning and another is unsupervised learning. The performance of the supervised learning depends on the set of training data. It does not ably perform well without abundant data set. Supervised learning method teach a sentiment classifier based on training data set. The classifier is used to identify different sentiment on each attribute. There are many learning based classification models are available. SVM, Naive Bayes, and Maximum Entropy (ME) model these are the learning based classification model [5].

3.3.1 Categorizing Positive and Negative Tweets

If the total score (+ve score, -ve score) is greater than 0, one can consider it as positive tweet and if it is smaller than 0, one can consider as negative tweet. In this approach, one can count the number of positive and negative tweets. But the positivity or negativity depends on their tweet score.

a. Positive tweets

b. Weak positive tweets

c. Strong positive tweets

d. Negative tweets

e. Weak negative tweets

f. Strong Negative tweets

3.3.2 Computing the Results

One can try to analyze the sentiments on the tweets and calculating as following results:

a. Number of tweets:
Count the number of tweets to be processed.

b. Total Number of positive tweets:
Let t is the token or words and sent(t) is the sentiment value. Then one can take the tweets as the +ve tweet if

$$\{(t \in Tweets)\&\&(sent(t) > 0)\}$$

c. Total Number of negative tweets:
Let t is the token or words and sent(t) is the sentiment value. Then one can take the tweets as the -ve tweet if

$$\{(t \in Tweets)\&\&(sent(t) < 0)\}$$

d. Weighted mean:
Weighted Mean =

$$\left(\sum_{i=0}^{\pi} wiai\right) \div \left(\sum_{i=0}^{\pi} wi\right)$$

Here, wi =weight and ai = value of the tweet. If one want to give important values to the main attribute of any product or service, one could use weighted mean. It is given priority focus to the main characteristics more than other usual characteristics.

e. Arithmetic Mean:
Arithmetic Mean = (Total value of the tweets÷Total no. of tweets)

Arithmetic Mean is given equal importance to all data and this is the basic difference between weighted and the arithmetic mean.

f. Positive Sentiment by Percentage:
PS(%) = (Number of positive sentiments ÷ Total no. of tweets) ×100

g. Positive Sentiment by Percentage:
PS(%) = (Number of negative sentiments÷ Total no. of tweets)×100

Our analysis could give such results as in the following diagram:

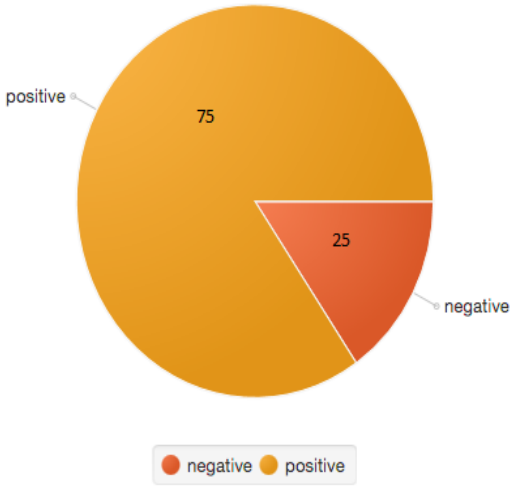

Fig 3.1: Pie chart of Tweets (In Number)

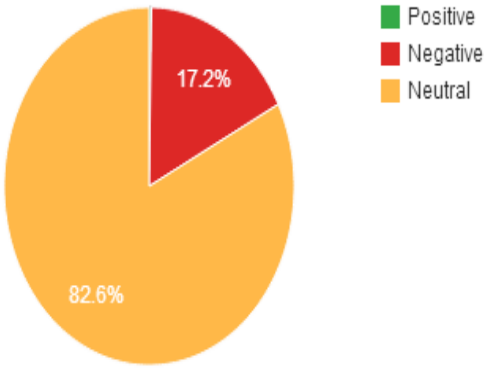

Fig 3.2: Pie chart of Tweets (In Percentage)

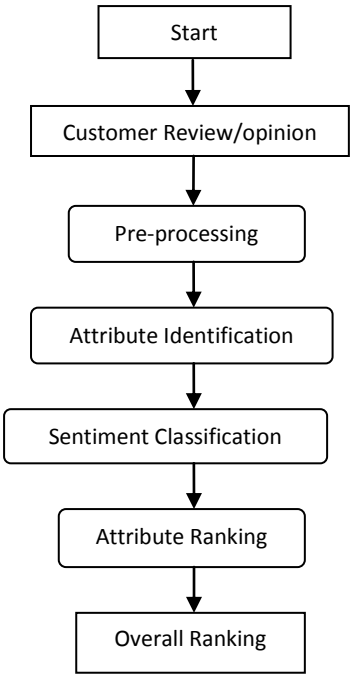

Fig 3.3: Flow Chart

3.4 The probabilistic ranking algorithm for the product ranking

In the proposed framework the important attribute that are frequently mentioned in the review/opinion have a great significance on the product rating. The customers review important attributes frequently. This frequent review/opinion about the important attributes greatly affects the overall opinion on the product. The final opinion in a review by the user is an aggregate of the opinions given to specific attribute or the feature of the product in the review, and various features have different contributions in the aggregation i.e. the sentiment on un-important or less important attributes and important attributes have impacts on the generation of the overall opinion. In this paper one can propose to use AFINN dictionary version111 approach. AFINN is a list of English words rated valence with an integer between minus five (negative) and plus five (positive) [9].

3.5 Pseudo code algorithm

- Start Program
- Locate the Comments/Feedback section within the website
- If comments/feedback != 0
  - Do Until the last comment
    - Read the first comment
      - Tokenize the sentences in the comments.
      - Identify the attribute/s from the sentence.
      - Give the weightage to the attribute if not already assigned
      - Calculate the frequency of the attribute.
      - Alter the weightage of the attribute if necessary
      - Analyse the sentiment of the sentence.
  - End Do
  - Based on the weightages of all or important attributes mentioned in the comments assign the overall rating to the product
- End of the Program

4. APPLICATIONS

The attribute ranking framework is very useful in a large number of real world applications. The applications of the sentiment analysis are endless. It can be used in social media monitoring, capturing the Voice of Customer (VoC) to keep updated or to track the customer review/opinions, survey responses etc. Businesses and organizations are interested in review/opinions to use them for new product conception, brand idea, brand management, product and service benchmarking, market intelligence etc. Whereas other users can use the review/opinion of others and take into account the ranking while purchasing a product or the services.

5. CONCLUSION

For this paper authors have surveyed reference papers relating to attribute identification, sentiment analysis and sentiment classification. Author's theory is that the key attributes of the product are the ones which are more frequently reviewed/commented by the users and the customers. These key attributes carry a lot of weight on the overall review/opinion on the product. Based on the theory, one can try to develop an attribute ranking algorithm which will identify the important attributes by simultaneously considering the attribute frequency and the stress that the consumers' feedback given to each and every attribute/feature on their overall reviews/opinions. In the future thepapers work can be expanded in many ways including,

1. It can be expanded to target a different domains ex. Telecom, Banking, Sales, Marketing etc.

2. Alternatively the paper can be slightly changed to do the sentiment analysis from different websites. Ex. e-commerce websites, other social media websites like facebook etc.

3. Another future change is to make it analyze the sentiments using different International and regional languages.

6. REFERENCES

[1] W. Jin and H. H. Ho, "A novel lexicalized HMMbased learning framework for web opinion mining," in Proc. 26th Annu. ICML,Montreal, QC, Canada, 2009, pp. 465–472

[2] Bing Liu, "Sentiment Analysis and Opinion Mining" pp.7-140,2012.

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. SIGKDD, Seattle, WA, USA, 2004, pp. 168–177.

[4] A. Finn, Kushmerick, N., and Smyth, B. 2002. Genre Classification and Domain Transfer for Information Filtering. In Proc. of European Colloquium on Information Retrieval Research, pages 353-362.

[5] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," J. Emerg. Technol. Web Intell., vol. 2, no. 3, pp. 258–268, 2010.

[6] T. L. Wong and W. Lam, "Hot item mining and summarization from multiple auction web sites," in Proc. 5th IEEE ICDM, Washington, DC, USA, 2005, pp. 797–800.

[7] Chetan Mate, "Product Aspect Ranking using Sentiment Analysis: A Survey"in IRJET vol.03, Issue 01, pp 126-127.

[8] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in Proc. 14th Int. Conf. WWW, Chiba, Japan, 2005, pp. 342–351.

[9] Finn Årup Nielsen, March 2011 http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[10] Bing Liu, "Sentiment Analysis and Subjectivity" Department of Computer Science, University of Illinois at Chicago, Handbook of Natural Language Processing, Second Edition

[11] Md. Ansarul Haque, Tamjid Rahman "Sentiment Analysis by Using Fuzzy Logic" Department of Computer Science and Engineering, Stamford University, Bangladesh. IJCSEIT, Vol 4_No, 1, February, 2014