Extracting and Ranking Aspects from Customer reviews using Natural Language Processing Techniques

B.Rayudu[1], Dr. P. Sujatha[2]
[1]*Student, Department of Computer Science, Pondicherry University, Pondicherry, India*
[2]*Assistance Professor, Department of Computer Science, Pondicherry University, Pondicherry, India*

Abstract — **Now-a-days, shopper's reviews are playing the most significant role on the society not only for customers but also production companies. A huge amount of customer reviews are available everywhere. These reviews are terribly useful to induce quality information about concern product and its aspects. Here Aspects are the features or components or attributes of a service or product. Aspects play the main role for products sometimes it can decide the product performances and its performance may impact on the product also. This article is proposed Aspect ranking model consisting of three phases i) Extract product aspects ii) Identify aspect sentiments iii) Aspects ranking. This model is useful for both customers and firms. The datasets used for this model are SemEval2014 restaurant, SemEval2014 laptop, Amazon canon G3 camera. The proposed model compared with the counter parts and achieves accurate results.**

Keywords— Customer reviews, Aspect Extraction, Sentiment Analysis, Aspect based Sentiment Analysis, Product Aspects, Aspect Polarities, Abstract Syntax Tree, Pos Tag, Bag of words, Aspect Ranking

I. INTRODUCTION

E-trade has attained abundant recognition in the last couple of years. Humans are showing their interest in online buying [1] because it is more convenient, time-saving and smooth way. There are loads of heaps of merchandise offered online by means of numerous traders and lots of retail websites offer a platform to promote their products online. The websites like Flipkart, Amazon and Paytm are popular to promote their merchandise online [2]. These websites provide a provider to submit feedback about products. Customer reviews contain important information like the standards and quality of the products but sometimes customer reviews are very difficult to understand that time product aspects play the main role. Aspects are the features or components or attributes of a service or product. For example, one review on a laptop is "*Battery of HP laptop is bad*", right here which exhibits a negative opinion on the aspect "*Battery*" of HP laptop [2]. Each and every product may contain many aspects. Some aspects are very important to other aspects of product. Those essential aspects have a mile's greater impact on making decisions of consumers and the product improvement method of the industries. As an example, the Samsung mobile product has aspects like "*camera quality*", "*screen*" and "*battery*" here more vital is *screen* [1]. According to the corporations, the important aspects of the evaluations aren't best beneficial for product quality development however also helpful for reinforcing and enhancing their popularity. But these elements are not at once the general opinion of the consumer on a product and it can't change their purchasing decisions. For example, "*Redmi 4 has worst battery power*" still it is highly rated. The main reason behind it is that other aspects like "*design*" and "*usability*" may be good. Most of the previous works were based on text classification but not aspect wise mostly decided overall text summary but not individual aspects wise, and most of the works done using two or three pre-trained datasets [1][2] [7] if any new review is present then the system is not able to perform any functionalities.

A. What is Aspect Based Sentiment Analysis?

This is a text analysis technique that can classify data and then determine the sentiment assigned to it [10][11]. For analysing the customers comment and feedback by linking particular sentiments with various products plus service aspects ABSA [5][6] is used. Aspect means; services, products components or the attributes e.g., user experience, queries response time. Sentiment analysis is used for classifying each text polarity [9]. Review may be positive or may be negative and even contain both reviews. In text analysis text is broken down into aspects then allocates each and everything in sentiment level. This helps businesses in becoming customer centric, it involves listening to customers, accepting their voices, examining feedback, learning the customer experiences and understanding expectations of the customers for the produced products plus services. In sentiment analysis only sentiment of complete text is detected, and in aspect based every text is analysed for identifying various aspects and then determining equivalent sentiments for each one. Aspect-based sentiment

analysis will focus mostly on information behind the text, so results are even more accurate, interesting and detailed.

II. LITERATURE REVIEW

One majorly used technique in classification is the ensemble classification for finding the finest classifier for resolving all classification related troubles. Twitter sentiment classification's performance will gradually improve by ensemble classification [12]. SVM, Logistic Regression, Naive Bayes and Random Forest classifier are the various base learners by which ensemble classification is formed [11]. Results distinctly show performance of the ensemble classier is superior to stand-alone classifier. This helps companies in monitoring all consumer opinions about products plus for the consumers in choosing finest products depending on community opinions. One possibility ranking model will be usually used for identifying one necessary aspect regarding product based on the customer reviews. Pos tagging system is used to extract aspects, but pos tagging does not always give proper clear results due to some uncertainty and confusions like word 'awesome' will be known both as noun and adjective in this pos tagging [3][4]. Dominantly nouns are given as answers mostly and possibility for giving wrong results [1] is more. For predicting and calculating aspects probability aspect ranking [2] is typically used based on customer reviews.

Aspect ranking is a challenging task in natural language processing [5] finding aspects from customer reviews are easy task but assigning polarities to those aspects is really complicated task and after assigning polarities find ranking is also challenging task. In most of the existing works adjective considered as an aspect but it is false in my view because adjective always is talking about polarity not feature of product or service but proposed model removing conflicts successfully between adjective and noun terms easily. Most of the existing systems work on specific predefined data sets only, if any review occurs other than predefined dataset those systems which are failed to produce results. But this article proposed Aspect Ranking model it is shown in Fig. 1. It is successfully producing results accurately irrespective of data sets and reviews.

III. ASPECT RANKING MODEL

Aspect ranking model shown in Fig. 1 consist of five modules (1) *Preprocessing:* for pre-processing the data (2)*Pos Tagging:* for detecting speech of each token (3) *Aspect Extraction:* for finding aspects (4) *Aspect classification*: for classifying aspects (5) *Aspect ranking*: for ranking aspects, each module details are clearly presented below.

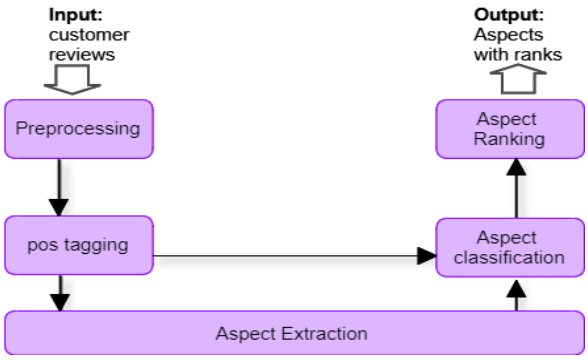

Fig. 1. Aspect ranking model

A. *Prepossessing the Dataset*

Converting any form of dataset into text file format only because of simple and easy access of data, data is arranged in line-by-line manner, for that 'punkitlearntokenizer' is used. Then after all the stopwords are removed here and finally data is converted *into* uppercase letters.

B. *Pos tagging*

Pos [3] tagging is used to identify the parts of speech in a comment. For parts of speech, data or comment should be tokenized only other than that not possible, pos tag assign the parts of speech tag to each and every token and indicating the speech with tag in Table 1 shows the all-possible tags with their speech description. In Fig. 2. Shows the pos tagging and tags indication of each token for the comment *"she sells seashells on the seashore"*.

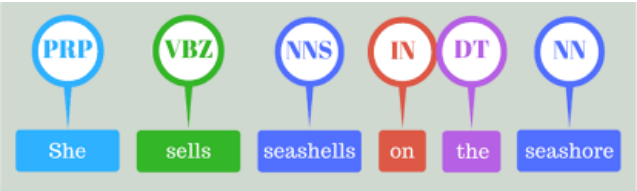

Fig. 2. Pos tagging with example

C. *Aspect Extraction*

For aspect identification purposes Pos Tag [4][8] with Bag Words method is used, in this method first predicting the all-noun phrases by using pos tags (NN, NNS, NNP, NNPS) method, next maintained a bag with all possible positive and negative words because most of the words are comes under noun [3] and adjective so, always checking every noun term with bag of word if it is presented in bag then it is not considered as aspect, for removing conflicts between noun and adjective bag of words method is used. For converting any

form (V1, V2, V3) of word into their root word lemmatization operation is applied to the noun terms then get perfect noun terms, those noun terms considered as Aspects of the product or service.

Table 1. Pos tags with Descriptions

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

D. Aspect Classification

For the purpose of classification [8] AST (Abstract syntax Tree) [5] is used. In Fig. 3 shows the AST structure and classifying the tokens based on parts of speech for the comment *"The food is top notch and restaurant atmosphere is nice"*. By using AST can find the polarity of aspects. Here root considered as sentence and branches as parts of speech. Noun term all are the aspects and adjective terms all are the polarity of aspect.

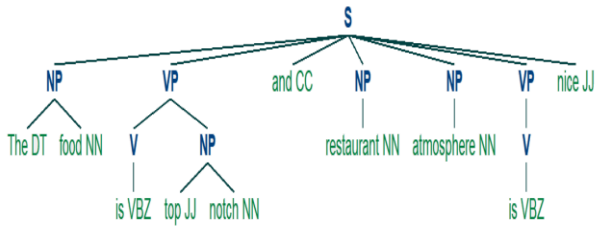

Fig. 3. Abstract syntax tree structures.

E. Ranking of Aspects

After completion of classification the data is stored in csv file format [1][2] [7]. Filtering the data in csv file and find the appropriate aspects that is top at most 25 only considering based on positive, negative polarity percentage, Aspect scores substituting in formula-1 it producing ranks of all the aspects accurately.

Range1 = (Max1–Min1) // (-100 to +100)

Range2 = (Max2 - Min2)//20 (0 to 19)

Value = (AspectTrue score - AspectFalse score)

Rank=(((value–Min1)*Range2)/Range1)+Min2 → formula-1

IV. ASPECT RANKING ALGORITHM

Initialize T, R, P as Tokenizer, stop words remover, pos tagging
Initialize P1, N1 as positive words list, negative words list
Initialize A1, A2 as AspectTrue, AspectFalse
Initialize T1 as empty string
While comment in reviews:
Step1: ● Apply T, R, P to comment and store in T1
Step2: ● Extract aspect Terms from T1 using P
Step3: ● calculate polarity of T1 by using Abstract syntax tree
● for word in polarity:
● if word is adjective and equal to negative word
● Increment A2 by 1
● Else
● Increment A1 by 1
Step4: ● Substitute A1, A2 scores to formula 1

A. Aspect Ranking Algorithm explanation

Step 1: Accessing the comments from text file one by one. Each comment is filtered like removing stopwords and converts into uppercase letters. After filtering those comments assigned to tokenization and finally parts of speech tag is assigned to those tokenized comments one by one.

Step 2: For find aspects of a review, that is noun terms of a comment. Taking pos tagged data and extract the noun terms (NN, NNS, NNP, NNSP) from the comment.

Step 3: Assign the polarities to aspect, taking pos tagged data from step1 and construct an AST (abstract syntax tree), the tree is arranged in a manner like sentence, noun phrase, and verb phrase and adjective finally the leaf nodes are tokens only. Here deriving the polarities from adjectives and aspects from nouns. If adjective is equal to negative word, then it will be considered as aspect polarity is one like that increase the negative count, if not a negative word comes under adjective, it is absolutely a positive word so automatically assign count 1 to that aspect and increase by one to positive count. Finally, will get the aspect with negative percentage and positive percentage.

Step 4: For the purpose of ranking need to execute the formula 1, taking aspectfalse and aspecttrue that is aspect negative

percentage and positive percentage values from step 3 substitute in formula 1 with this will get ranks of all aspects.

V. EXPERIMENT AND RESULTS ANALYSIS

A. Datasets

The data sets Collected from different websites which are SemEval2014, semEval2016, amazon and kaggle. SemEval, kaggle data sets are in XML, HTML format. Extracted product feedbacks from XML, HTML files those feedbacks are arranged in line-by-line order and stored in a text file. Three datasets are used in this model which are restaurant, laptop, canon G3 camera, the restaurant, laptop datasets are gathered from SemEval2014 websites and canon G3 camera data set gathered from amazon.

B. Implementation Details

For implementing this research, windows 10 operating system with 8GB RAM and 1 TB ROM are used, for coding purpose python programming with python 3.7 IDLE tool, NLTK toolkit [4][1] [5]is used. NLTK toolkit containing nearly 300 libraries and separately installed pip, numpy, pandas etc... For implementing mathematical expressions mathplot, GaussianNB and sklearn packages are used.

C. Results Analysis

Fig. 4. Shows the results with SemEval 2014 Restaurant dataset, here high rated aspects are 24 only. In that '*wine*' is 40 percent positive opinion and 7 percent negative opinion among all *wine* is top rank due to more positive opining and less negative opinion. '*Sushi*' is least rank due to more negative opinion and less positive opinion. Fig. 5. Shows the results with SemEval2014 Laptop dataset here '*lightweight*' is top rank due to more positive opinion and less negative opinion but '*call*' is least rank due to more negative opinion less positive opinion. Fig. 6. Shows the results with Amazon canon G3 camera dataset here '*Research*' is top rank and '*SLR*' is least rank due to their positive and negative opinion

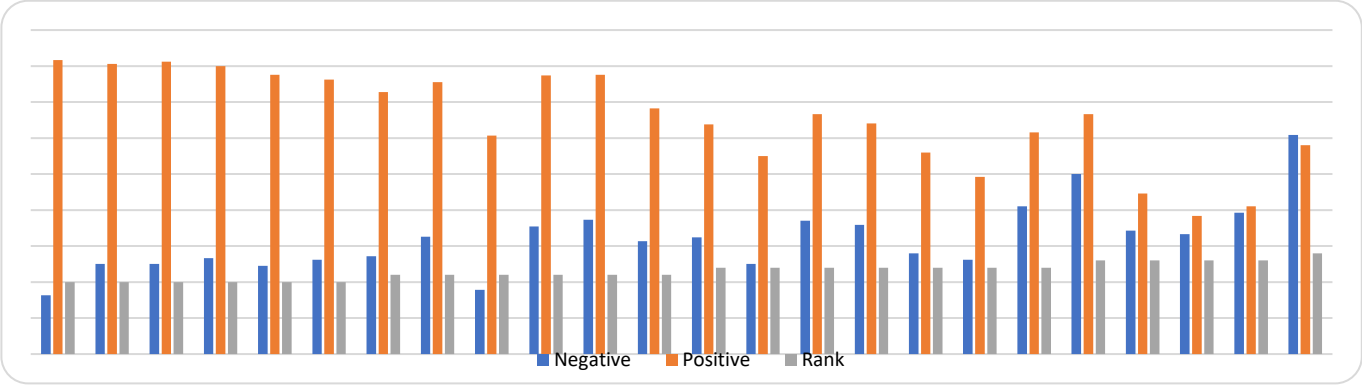

Fig. 4. Results with SemEval2014 Restaurant dataset

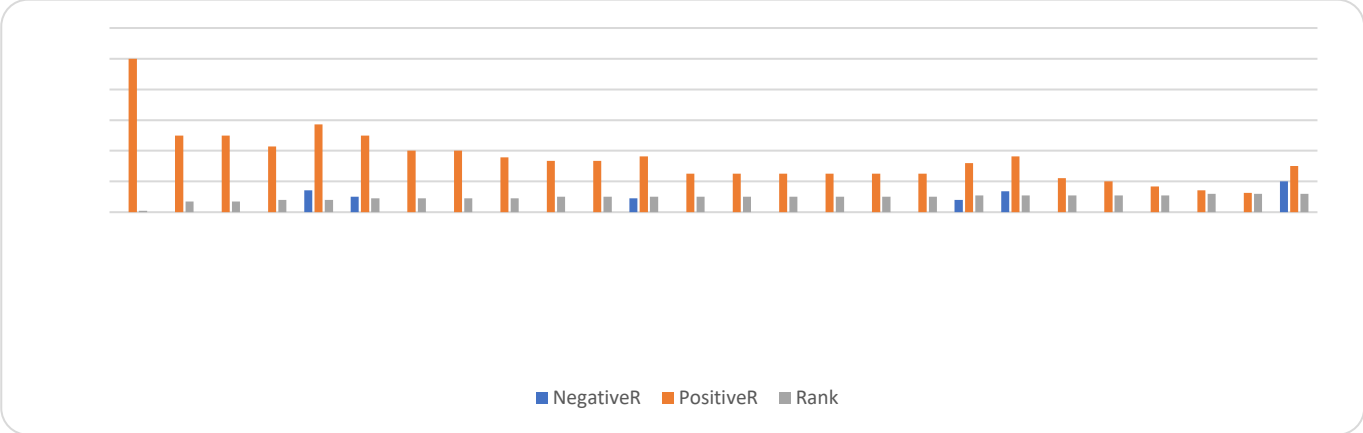

Fig. 5. Results with SemEval2014 Laptop dataset

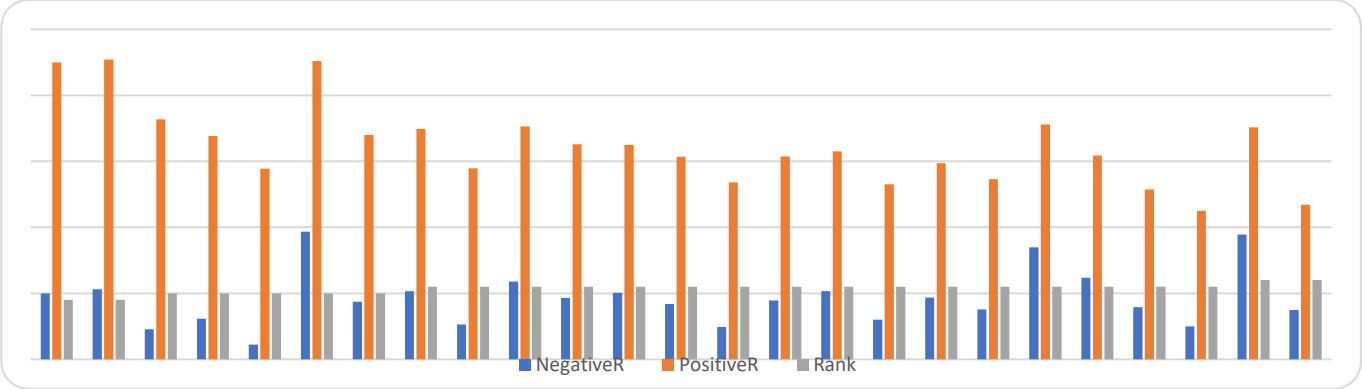

Fig. 6. Results with Amazon canon G3 camera dataset

VI. CONCLUSION

The main objective of this project is to provide a model to customers as well as firms to predict the important aspects and aspect ranks. The Overall model works well with providing accurate results. Future enhancement of this model is to build a module to divide the sentiment into opinions. This model will work only comment consisting of one opinion if comment consists of more than one opinion then it will not produce accurate results. It is more confusing to assign polarities to the aspects.

REFERENCES

[1] J. T. M. W. Zheng-Jun Zha, Jianxing Yu and T.-S. Chua, "Product aspect ranking ad its applications," *IEEE Transaction on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 1211–1222, 2014.

[2] P.M. K.RutujaTikait, Prof.RanjanaBadre, "Product aspect identification and ranking system," *International Journal of Science, Engineering and Technology Research (IJSETR)*, no. 4, pp. 304–311, April. 2015.

[3] L.F.FeilongTang, "Aspect based ne-grained sentiment analysis for online reviews," *Elsevier*, pp. 190–204, 2019.

[4] J. Thanaki, "Python Natural Language Processing." *Packt Publishing Ltd*, Birmingham, UK, pp. 1- 486, 2017. [Online]. Available: https://www.packtpub. com/product/python-natural-language-processing/9781787121423.

[5] Federico Pascual,"A comprehensive guide to aspect-based sentiment analysis" *monkylearn",* 2019 [Online]. Available: https://smonkeylearn.com/blog/ aspect-based-sentiment-analysis/.

[6] Naveen Kumar Laskari , Suresh Kumar Sanampudi, "Aspect Based Sentiment Analysis Survey", *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 18, pp. 24-28, march-april. 2016.

[7] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, "Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews," in Proceedings of the 49th Annual Meeting of the *Association for Computational Linguistics: Human Language Technologies* (ACL 2011). ACL, pp. 1496–1505. 2011.

[8] Zhiqiang, Toh, and Wang Wenting, "DLIREC: Aspect Term Extraction and Term Polarity Classification System.", *8th International workshop on Semantic Evaluation (SemEval2014*), pp. 235-240, Aug. 2014.

[9] Gupta, De1epak Kumar, and Asif Ekbal, "IITP: Supervised Machine Learning for Aspect based Sentiment Analysis.", *8th International workshop on Semantic Evaluation (SemEval2014*), pp. 319-323, Aug. 2014.

[10] Pontiki, Maria, et al., "Semeval-2014 task 4: Aspect based sentiment analysis.", *8th International workshop on semantic evaluation, SemEval 2014*, pp. 27-35, Aug. 2014.

[11] Brychcın, Tomáš, Michal Konkol, and Josef Steinberger, "UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis", *8th International workshop on semantic evaluation, (SemEval 2014),* pp .817-822, Aug. 2014

[12] Ankita, NabizathSaleena, "An Ensemble classification system for twitter sentiment Analysis", *international conference on computational intelligence and data science*, pp. 937-946. 2018.