RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

HYBRID APPROACH FOR ROOT CAUSE ANALYSIS OF ASPECTS USING ONLINE REVIEWS

Blessy Selvam[*1], Dr. S. Ravimaran[2] & Dr. Sheba Selvam[3]
[*1&2]Dept of CSE ,Anna university (Bit- Campus), Trichy, India
[3]Dept of CSE, Rajarajeswari College of Engineering, Bangalore, India

ABSTRACT

What they may think" has always beenpredominant for most of us during the process of decision making. Today user opinion on products is increasing day by day on the web. But, in point of fact it is difficult for a customer to examine the reviews to decide to buy the product or not. Also the producer of the product needs to govern the opinion of the customer. Identifying the aspects corresponding to a target and its sentiment is a crucial part in opinion mining. This work proposes machine learning methods for identifying implicit and explicit aspects. The aspects are categorized as positive and negative using Support Vector Machine (SVM). Finally, the classified opinions are further grouped into aspect categories using K- means clustering, which allows the identification of positive and negative root causes for a product. Experiment conducted on benchmark review data set which includes reviews about five distinct products. The Performance achieved by the proposed hybrid approach is compared with CNN+ LP, POPSCU and TF-RBM state-of-the-art models. The results of the proposed method shows improved Precision, Recall and F1-score that outshine the state-of-the-art baselines.

Keywords: *Opinion mining, aspect extraction, sentiment analysis, Support vector machine, K-Means clustering.*

I. INTRODUCTION

Opinion mining has tempted huge amount of recognition from researchers of natural language processing in the past years because of the existence of many research problems and empirical appeal. The sentiment obtained from product reviews is beneficial to customers and companies. It is helpful for the companies to improve their products or services from the reviews posted on the Web [1]. The Websites yield an extensive source of consumer reviews, but it is hard to read every one of the review in order to obtain upright estimation of a product [2]. Spotting out the defects in a product is effective for companies to enhance their fierceness situation. Methods were developed to help companies to benefit from these resources to know their business status in the real world environment. Opinion mining is concerned with analysis of the opinion of a target. Finally, the highly weighted opinions about a certain concerned product are prescribed to the end user. For promoting business big companies and business officials are giving privilege for opinion mining.

Aspect-based opinion mining looks at the opinions and performs fine grained analysis of reviews. There are many aspects that a customer would look for a restaurant to dine in, such as food, service, and ambience. Among these aspects [3], the type of food and quality of food are the most important. Therefore, the restaurants are rated based on customer reviews. An aspect that appears as noun or noun phrase in a review is known as the explicit aspect, for example, "food taste was good", food is the entity, and taste is its explicit aspect. Whereas, "good" is the opinion expressed over that target entity. Implicit features are the features which are not evident in a review. For example [4], in review "The hotel is expensive", the aspect referred is "price" although word "price" is not mentioned abruptly.

II. RELATED WORK

The major of research work on opinion mining has been dealt with using the English language, as English is the dominant language of scientific researchers. Below discussed are some of the existing works in sentiment analysis.

In [5], Conditional Random Field (CRF) method has used probability based learning to extract the aspects. The product aspects are divided into three categories. After aspect extraction, two similarity calculations were made to

classify the aspects using wordNet.A supervised approach was discussed in to determine the aspects which are implicit from review dataset. Implicit aspects were generated and scored based on the co-occurrence and frequency of the each word[6]

A threshold is set for identifying the aspects that are implicit based on the frequency of occurrence.

Minimum support threshold finds the frequent aspects for a review. Naive Bayes [7] identified the sentences as positive or negative using term counting approach. In [9] the aspect and polarity of products is sort listed based on the weightage of user's interest.

A dependency pattern-based approach based on bootstrapping extracts aspects along with its polarity. After aspect extraction the semantic similarity based technique [8] groups the similar aspects and calculates the position of each aspect sort out the significance of the aspects. So that it is feasible to understand and estimate the frequency user comments on each and every aspect.

Semi supervised alignment model is proposed in [9], here the relations of the opinions are identified. The confidence of a candidate is estimated based on co-ranking approach and represented using a graph. Highly confident opinion words are extracted. Opinion relations are found more accurately using this model than nearest-neighbour.

Feature selection method selects subset of features that are informative for improving the prediction results. An unsupervised approach namely k- means in [10] examines the generated features subset quality. The hybrid particle swarm optimization algorithm (H-FSPSOTC) shows improved results when compared with other models.

Presented entity-level opinion mining in [11] using lexicon technique shows high Recall and low precision. An automatic identification of opinion from twitter data is followed to improve recall. A classifier is used to find the polarity. The tweets were assigned polarity based on the retrieved opinions.

An unsupervised multilingual syntax-based rule was given in to perform well on different corpa and domains. Based on Part-of-speech tagged and dependency parsed aspects were extracted[12]. It addresses semantic composition and it outperformed other supervised methods.

In[13], a novel algorithm called "aspectator" detects and rates aspects from reviews automatically. The "aspectator" pairs the dependency path syntactic between different words. Ten handcraft dependency paths were outlined to discover product aspects and opinion words. Initially "aspectator" mines opinion pair from the generated syntactic dependency tree. The whole opinion set is explored with the opinion pair identified by observing with neighbouring words. WordNet and Senti-Wordnet were handled to cluster and finding the polarity of the opinions.

Cross domain relevance for product aspects [14] firstly extracts the pairs containing the aspect and opinion set. The Intrinsic and Extrinsic domain relevance aspects were scored. The aspect holding opinions were found using the intrinsic and extrinsic domain relevance value. It is found that the intrinsic values are greater when compared with the extrinsic values based on threshold values.

A hybrid approach for finding product aspects was given. The substring, dependency and constrained topic model rules are appended with the association rules for identifying product aspects.
The WAM (word alignment model) and PSWAM (partially supervised alignment model) both with small and large datasets were used. They tested and found that WAMperforms well for small data set and fails for large datasets. And hence they have proved experimentally that PSWAM outperformed WAM both on small and large datasets.

In the above existing papers only limited work has been done for identifying explicit and implicit aspect and also the performance level obtained in the existing methods was low. So, the proposed work deals with the above problems.

ISSN 2348 – 8034
Impact Factor- 5.070

III. PROPOSED WORK

The proposed hybrid model as given in Fig. 1, contributes the following:

- Pre-processing of reviews
- Extraction of aspects
- Mapping of word vector into numerical vector using word embedding
- Classifying the aspects using SVM
- Aspect categorization using K- means clustering
- Ranking and user feedback of aspects

The proposed hybrid approach extracts opinion of aspects from online reviews. After aspect extraction the sentiment of the aspects is found. The number of opinions containing both positive and negative opinions corresponding to each aspect is scored based on frequency count. And the final prediction is given as feedback.

A. Preprocessing

The initial step in the proposed work is pre-processing as the opinions are from online reviews. An online review is given for pre-processing. At this stage Stop words are removed from the reviews because these are words which carry no information. For example, Words like is, are, was etc. are removed. Stemming eliminates variation in words, such as “computer”, “computing” in order to represent each word in its general form.

B. Aspect extraction

Tokenization breaks a stream of text content into a single word, terms or some other significant part called tokens. The review data is given as input which is normally a set of reviews stated by the customer regarding a product or an organization etc. During the aspect extraction phase the reviews are divided into tokens namely aspects, which can be explicit or an implicit aspect. For example "The phone quality is good.” Here the aspect “phone quality” is mentioned directly in the review, hence it is explicit aspect. For example, “The camera is small enough to put in my pocket.", the aspect "size" is not directly mentioned but instead it is implied as “small” which is nothing but the “size” of the entity camera.

Aspects are recognized by examining the POS component in the review. The Part-of-speech of a word is determined and extracted in a syntactic manner. The aspects are finally extracted and given for frequency based word embedding process.

C. Word Embedding

Word embedding is as representation of scatter of words, which computes word vector based syntactic and semantic format. The word vector representation is dense. Hence, it is converted to low-dimensional numerical vector. Word embedding is the process of mapping the given word vector with vectors of numbers. Word embedding is classified into frequency based and prediction based models. Frequency based model examines the number of occurrence of each aspect. By using the frequency based model a numerical vector is generated. This numerical vector is given as input to the SVM classifier for further prediction

ISSN 2348 – 8034
Impact Factor- 5.070

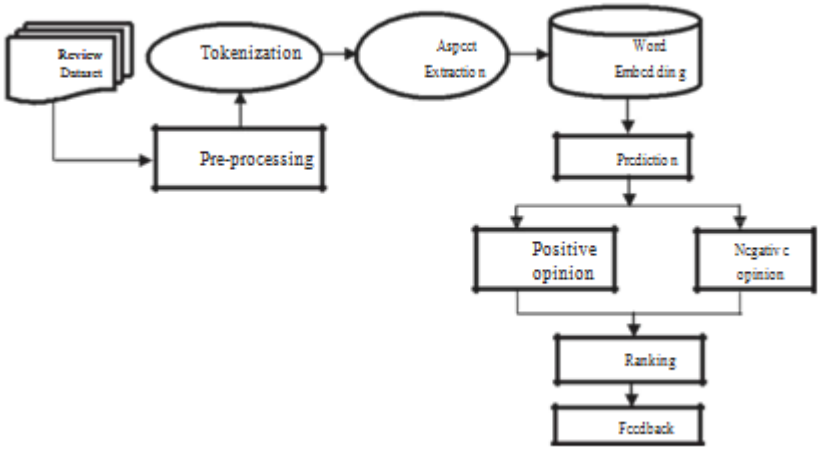

Fig.1. Proposed Hybrid Approach for Root Cause Analysis

1 . Hybrid Approach

The proposed framework classifies the extracted aspects as positive and negative. For example "The best thing I tasted were the lamb hops" in the above review sentence it stated the aspect "lamb hops" which comes under the food category and the opinion expressed for that aspect is positive. For classification the proposed hybrid approach worked on supervised method SVM. This supervised approach is shown to be one of the best supervised machine learning algorithms in many opinion mining research work and is also mentioned in these work [15] [16] [17]. These papers suggest that SVM classifier achieved improved performance in their work.

The Support Vector Machine defines a linear hyper plane which separates the aspects such that the margin is maximized. The logic behind SVM is that once the decision boundary (margin) is maximum then the misclassification rate is minimum. The hyper plane is a line dividing a plane in two parts in a two dimensional space. Where, each part represents two different classes. Once, the dissociation of the margin is larger, then it is said to be a good margin. The points will not cross each other in a good margin and lets the points to occupy their respective classes. Once, SVM has classified the aspects as positive and negative, the similarity between two words is measured using lexical similarity [20] for grouping synonymous aspects. For example, "phone" and "mobile" have greater similarity in WordNet. The WordNet lexical similarity intensifies the strength of similar words [24]. The highly similar words are synonyms. Similarities are used to merge similar aspects.

2. The similarity between aspects are found using word Net

The identified similar aspect pairs are sorted in descending order
Top N pairs are considered and the pairs are merged.

Opinionated similar aspects are grouped into a single entity Known as Clustering. Clustering is method which groups the Dataset into clusters of smaller unit based on similarity [18]. One Of the algorithms used is k-means clustering.

The k-means algorithm is as follows, 1. Centroids "c" is selected arbitrarily
A. Closeness from each centroid "d" to all aspects "a" is calculated repeatedly.
B. Each identified aspects "a" with minimum distance is assigned to the centroid.
C. New centroids are calculated from each new data point

- Closeness is re-measured from each new centroid to every aspect a repeatedly till no aspects are found.

Firstly it determines k centre's arbitrarily for each cluster, the

Proposed work considered K=3 clusters for (aspect, cost and appearance) aspect categorization. Secondly, the distance between the aspects in the dataset is calculated and the centroid for each aspect is found[19]. The discovered aspects with minimum distance are assigned to the closest cluster. Sub-clusters are formed based on similar aspect category.

The Euclidian determines the distance measure between the aspects and cluster centroid. Let ***xi*** and ***yi****be two aspects.* Then distance between aspects is measured using the formula as given in "Eq. (1)".

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i, y_i)^2} \qquad (1)$$

For example (1)

"Service was quick" polarity="positive" Aspect term="Service" polarity="positive" Aspect category="service" polarity="positive"

In example 1, its aspect category is "service". Each sub cluster represents each aspect category, so aspect category is obtained for positive as well as for negative opinions based on similarity and distance measures.

F. Ranking

After clustering, the aspects within each cluster are ranked based on the frequency of appearance of the aspect terms [8]. Highly frequent aspects in both positive and negative classifiers after aspect categorization are considered and they are provided as feedback to the customer. Ranking of aspects is purely based on frequency of appearance. Implicit aspects are identified because of clustering and lexical similarity. As the explicit aspects are directly mentioned in the reviews.

IV. EXPERIMENTAL RESULTS AND COMPARISON

For experimental analysis benchmark dataset has been taken for consideration. The dataset [21] information is shown in table I.

The dataset have been annotated with aspects and sentiments. The aspect category included in this work is aspect, cost and appearance.

Experiment conducted with this dataset is compared with state-of-art models [22] Popescu, CNN+LP [23] and TF-RBM

C. These methods along with the proposed method are compared with metrics namely Precision, Recall and F1-score.

$$\text{Precision} = \frac{tp}{tp + fp} \qquad (2)$$

In "Eq. (2)" tp refers to correctly detected aspects; fp denotes falsely detected aspects

$$\text{Recall} = \frac{tp}{tp + fn} \qquad (3)$$

In "Eq. (3)" tp refers to correctly detected aspects; fn refers to missed aspects.

ISSN 2348 – 8034
Impact Factor- 5.070

TABLE I: Dataset Description

Data	Total Reviews	Reviews with Polarity	Reviews without Polarity	Aspects Featured	% Featured Aspects
Canon Digital Camera	597	238	359	237	40%
Nikon Digital Camera	346	160	186	174	50%
Nokia Phone	546	265	282	302	55%
Creative Mp3 Player	1716	720	996	674	39%
Apex DVD Player	740	344	396	296	40%

The Precision comparison shows that the proposed hybrid approach gives improved Precision values when compared with other state of art models. But,shows slight variation in Mp3 and Apex DVD products because of the absence of polarity for some of the reviews. Comparatively the Precision values for the first three dataset is good as shown in table II.

TABLE II: Precision Comparison

Technique / Data	CNN+LP	POPSCU	TF-RBM	SVM+K-Means
Canon	0.93	0.89	0.8	0.94
Nikon	0.82	0.87	0.87	0.89
Nokia	0.9	0.89	0.92	0.92
Mp3	0.92	0.86	0.86	0.83
Apex DVD	0.93	0.9	0.88	0.83

The Recall comparison manifest the hybrid approach gives improved Recall values when compared with other models. The proposed approach shows high Recall values for all five product reviews. This is because of accurate data selection. Recall comparison table is shown in table III.

ISSN 2348 – 8034
Impact Factor- 5.070

TABLE III: Recall Comparison

Technique / Data	CNN+LP	POPSCU	TF-RBM	SVM+K-MEANS
Canon	0.85	0.8	0.89	0.91
Nikon	0.87	0.74	0.93	0.94
Nokia	0.84	0.74	0.93	0.95
Mp3	0.86	0.8	0.93	0.94
Apex DVD	0.88	0.78	0.9	0.92

The F1 Score measures the accuracy of the hybrid approach. F1-score is a harmonic average of recall and precision values shown in Eq (4). Comparison shows that the proposed approach gives improved F1 Score values when compared with other state of art models. But for MP3 and Apex DVD the F1-score values are lightly less.F1 Score shown in table IV.

$$F1-Score = \frac{2(Precision * Recall)}{Precision + Recall} \quad (4)$$

TABLE IV: F1 Score Comparison

Technique / Data	CNN+LP	POPSCU	TF-RBM	SVM+K-MEANS
Canon	0.88	0.84	0.84	0.92
Nikon	0.84	0.8	0.9	0.91
Nokia	0.87	0.81	0.92	0.93
Mp3	0.89	0.83	0.9	0.88
Apex DVD	0.9	0.84	0.89	0.87

The average values of the measures Precision,Recall and F1-score of the proposed hybrid and state of art model is given Fig. 2 . The proposed model outshines the existing models based on the above mentioned measures.

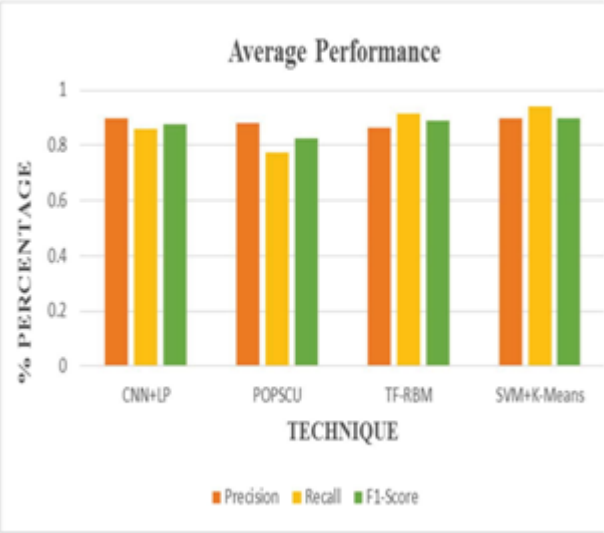

Fig. 2. Average performance of proposed and state-of-art models

V. CONCLUSION

The manufactures demand for a detailed fine-grained analysis of the piece of information expressed in a given review. Aspect-based opinion mining objective is to extract and classify the tenderness and opinion on a specific aspect or any analysis based on the interest of the user. Opinion mining is a major area which directly deals in analysing the Voice of Customer. This work dealt with identifying explicit and implicit aspect using the proposed hybrid model. The boon of the proposed hybrid model when compared with the state of art models is that the aspect word2vec conversion and the effectual recognition of implicit and explicit aspects. The categorization of aspects supports for improved prediction levels than the compared approaches.The proposed hybrid approach shows better Precision, Recall and F1 -score than the existing methods. The performance level of the proposed method is found to be good based on SemiEval dataset. In future, some other approaches can be combined for enhanced results. Finally, the identified aspects are ranked and provided as feedback to the customers. This work can be extended for mining opinions in comparative and sarcasm reviews.

REFERENCES

1. *Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A. S., ...&Hoste, V. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 19-30).*
2. *Schouten, K., van der Weijde, O., Frasincar, F., & Dekker, R. (2018). Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data.IEEE transactions on cybernetics, 48(4), 1263-1275.*
3. *Asghar, M. Z., Khan, A., Ahmad, S., &Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. Journal of Basic and Applied Scientific Research, 4(3), 181-186.*
4. *Prakhash, S., Nazick, A., Panchendrarajan, R., Brunthavan, M., Ranathunga, S., &Pemasiri, A. (2016, April). Categorizing food names in restaurant reviews. In Moratuwa Engineering Research Conference (MERCon), 2016 (pp. 1-5). IEEE.*
5. *Schouten, K., &Frasincar, F. (2014, July). Finding implicit features in consumer reviews for sentiment analysis.In International Conference on Web Engineering (pp. 130-144).Springer, Cham.*
6. *Jeyapriya, A., &Selvi, C. K. (2015, February). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on (pp. 548-552). IEEE.*
7. *Li, Y., Qin, Z., Xu, W., &Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. Multimedia Tools and Applications, 74(23), 10177-10194.*
8. *Li, Y., Wang, H., Qin, Z., Xu, W., &Guo, J. (2014, August). Confidence estimation and reputation analysis in aspect extraction.In Pattern Recognition (ICPR), 2014 22nd International Conference on (pp. 3612-3617).IEEE.*
9. *Liu, K., Xu, L., & Zhao, J. (2015). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. IEEE Transactions on knowledge and data engineering, 27(3), 636-650.*
10. *Abualigah, L. M., &Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. The Journal of Supercomputing, 73(11), 4773-4795.*
11. *Khan, A. Z., Atique, M., &Thakare, V. M. (2015). Combining lexicon-based and learning-based methods forTwitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89.*
12. *Bancken, W., Alfarone, D., & Davis, J. (2014, August). Automatically detecting and rating product aspects from textual customer reviews. In Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD (pp. 1-16).*
13. *Hai, Z., Chang, K., Kim, J. J., & Yang, C. C. (2014). Identifying features in opinion mining via intrinsic and extrinsic domain relevance. IEEE Transactions on Knowledge and Data Engineering, 26(3), 623-634.*
14. *Wang, W., Xu, H., & Wan, W. (2013). Implicit feature identification via hybrid association rule mining. Expert Systems with Applications, 40(9), 3518-3531.*

15. *Kai Yang, Yi Cai, Dongping Huang, Jingnan Li, Zikai Zhou, Xue Lei "An Effective Hybrid Model for Opinion Mining and Sentiment Analysis" 978-1-5090-3015-6/17/$31.00 ©2017 IEEE*
16. *Manoj Kumar Das, BinayakPadhy and Brojo Kishore Mishra "Opinion Mining and Sentiment Classification: A Review" International Conference on Inventive Systems and Control (ICISC-2017).*
17. *Wang, Gang, DaqingZheng, Shanlin Yang, and Jian Ma. "FCE-SVM: a new cluster based ensemble method for opinion mining from social media." Information Systems and e-Business Management (2017): 1-22.*
18. *p ratnababu and dr. Bhanuprakashbattula "a novel k-nearest neighbor distance based under sampling for improved opinion mining on skewed data using random forest" international journal of engineering & technology,2018.*
19. *Shoeb, Md, and Jawed Ahmed. "Sentiment Analysis and Classification of Tweets Using Data Mining." work4, no. 12 (2017).*
20. *Pedersen, Ted. "Information content measures of semantic similarity perform better without sense-tagged text." In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 329-332. Association for Computational Linguistics, 2010.*
21. *Rana, Toqir A., and Yu-N. Cheah."A two-fold rule-based model for aspect extraction." Expert Systems with Applications 89 (2017): 273-285.*
22. *Santosh, D. Teja, K. SudheerBabu, S. D. Prasad, and A. Vivekananda. "Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet." IJEME 6 (2016): 1-11.*
23. *Poria, Soujanya, Erik Cambria, and Alexander Gelbukh. "Aspect extraction for opinion mining with a deep convolutional neural network." Knowledge-Based Systems 108 (2016): 42-49.*
24. *Zhai, Zhongwu, Bing Liu, HuaXu, and PeifaJia. "Clustering product features for opinion mining." In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 347-354. ACM, 2011*