IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1464)
Available online at: https://www.ijariit.com

Opinion Mining for restaurant reviews using Naive Bayes Algorithm

Rutik Pravin Ambre
rutikambre.ra@gmail.com
Thakur College of Engineering and Technology, Mumbai, Maharashtra

Abhishek Sand
sand.abhishek@gmail.com
Thakur College of Engineering and Technology, Mumbai, Maharashtra

ABSTRACT

A wealth of unstructured opinion data exists online. Every day, millions of consumers add to this data when they share their opinion on a range of things, including feedback about their experiences with products and services. This feedback is volunteered, it contains the raw, unsolicited views and opinions about a brand, individual or event. Opinion Mining finds out the drivers behind the sentiment. By understanding what is driving the sentiment and how one is performing based on Net Sentiment, opinion data can be used to expose critical areas of strength and weakness. This data allows decision-makers in business, from customer experience and marketing to risk and compliance teams, to make the targeted, strategic overhauls needed to reinvigorate profitability or reclaim slipping market share. It is practically impossible to analyze all this reviews manually, so and automated aspect-based opinion mining approach is used. This paper focuses on aspect level, shows a comparative study amongst existing algorithm and proposes a new syntactic based approach which uses dictionary, aggregate score for opinion words. The dataset used was for restaurant reviews. The proposed method achieved a total accuracy of 87.06%.

Keywords— *Sentiment analysis, Prediction, Data Extraction, Data pre-processing, Opinion Mining, Parts of Speech*

1. INTRODUCTION

Opinion Mining is a field of study that analyzes people's opinions, attitudes, emotions, sentiment towards products, organizations, services. All of these comes under the branch of text mining and natural language processing. Opinion Mining is a sub problem of text classification in which texts are classified on the basis of emotions, opinion, aspects.

Due to tremendous growth in the social media, people now have opportunities to express their reviews through forums, discussions, comments, Twitter, etc. Users express their thoughts through reviews. These reviews are reviewed and analyzed by organizations and business, as they try to find out patterns and reasons behind the reviews and use this information for strategic planning and improved business intelligence. All of this is driving people more towards opinion mining.

The data today is vast (in zettabytes), it exists in both structured and unstructured form but the usefulness of data is increasing. There might be problems sometimes like inefficiency of machine learning algorithms to interpret this data. Hence to solve this problem model must interpret the aspects and emotions. Opinion Mining is the approach for it.

There are various levels for opinion mining that is sentence level, document level, aspect level. Sentence level involves fine-graining the sentence and identifying the different topics in document. And finally classifying it as positive, negative and neutral. In the case of restaurant reviews we require more fine-graining not only because it contains mixed emotions, opinions but also it consists of various aspects as well. Hence aspect-based opinion mining is suitable.

The key parts in aspect-based opinion mining are identifying the aspects, word orientation, orientation/polarity detection and opinion based on aspects. Consider an example of a review "The decor is superb but the service is worst as compared to decor." Here the aspects are decor and service and the opinion related to these aspects are superb and worse respectively. By analyzing this sample review, we found out that the polarity is positive for decor whereas negative for service this is a case of conflict in opinion. The current approaches provide the overall polarity for collective set of reviews but the proposed method goes deeper in intuition by consider the various aspects, syntactic dependency, parsing and also aggregated score for various aspects.

In the built model, we have taken reviews and collective score for various aspects and suggested the areas to work upon for improved business performance.

2. OBJECTIVES

The contributions of the presented work are as follows:

- To implement best approach available for detection of polarity of restaurant reviews using opinion mining techniques.
- To provide valuable insights by processing the reviews.
- Identify the various aspects, their polarity and orientation, and find the drivers behind the sentiment, and provide an aggregated score for each aspect

- Suggest areas for improvement by the aggregation of various score-based aspects.
- Visualize the findings from the processed dataset for improved business intelligence, marketing intelligence, help customers before visiting a restaurant.

3. RELATED WORK

One of the earliest works that were done in the field of aspects were extracting aspects using frequent itemset. The frequency of nouns and pronouns in the entire document were found out, then the Parts of Speech tagging for each word was given. After this it was sent to association rule mining for frequent itemset.

The next related work for finding aspects involved using machine learning algorithms. A labelled dataset with feature as expressions and target as aspects was used. SVM model was built for the dataset to find the semantic relationships between the expressions and the aspects.

The results of all the earlier methods were satisfactory but not good enough to be predict the aspect accurately, so since the development of natural language processing a lot of emphasis has been given on data cleaning. The cleaner the data the better it trains the model. By using methods like stopwords elimination, stemming, lemmatization and Count Vectorizer the data is processed in such a way that the semantic value is preserved and only the necessary words that contribute towards model building. If all the words are taken then the model becomes too complex and difficult to train. Once this data is processed it is fed to model like Naïve Bayes. Naïve Bayes algorithm works on probability, the higher the posterior probability of a particular aspect the review falls under that aspect. Naïve Bayes is a better classification algorithm as compared to SVM as it considers probability and not just dividing the data in particular classes.

As the data kept on increasing, various factors started contributing towards the aspect extraction like age, demographics even the length of the text. So the traditional document processing wasn’t enough. So researchers came up with even more fine grained approach for processing texts by going deeper into granularity. Rule-Based approach was used for opinion mining, the advantage this is rules are written in English language so generalized rules can be formulated using adjectives, adverbs, nouns. Once all these are extracted polarities of aspects were found out as positive, negative, neutral and conflict.

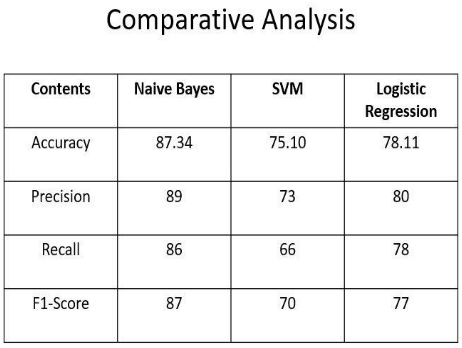
Further developments on this were made in the form of dependency parsing on the texts. Dependency parsing plays a pivotal role in semantic analysis which is a crucial part in natural language processing. Sometimes dependency parsing can be abstruse because of ambiguity. The most common structure for this is the parse tree. With the advent of libraries like Stanford Core NLP dependency parsing have been easier as it is available in both Java and Python and is faster.

Extracting aspect terms and opinionated texts are also possible using topic modelling approaches like Latent Dirichlet Allocation. The most recent developments in the field of opinion mining includes usage of deep learning concepts like LSTM and recurrent neural network. The classical RNN model involves only processing past information but bi-directional RNN involves processing both past and future information, this is specifically used in NLP as they provide more information for decision making, one way of constructing RNN is the stacking up of the layers.

Inorder to calculate the sentence level aspect score, heuristics are used which involves sending Adjective+Adverb and Adverb+Verb score using SentiWordNet. In these features are extracted using POS tagger and aspect terms are located then search is made for 5-10 gram forward and backward for relevant features.

4. PROPOSED METHODOLOGY

With our approach we aim to combine the aspects of various reviews and aggregate them to give an accuracy score for each aspect and based on the polarity of the aspect, areas for improvement will be suggested. Classification rules will be laid down, on the basis of the aspect polarity, semantic orientation will be taken to find out, whether the review is good or not.



Comparative Analysis

Contents	Naive Bayes	SVM	Logistic Regression
Accuracy	87.34	75.10	78.11
Precision	89	73	80
Recall	86	66	78
F1-Score	87	70	77

Figure 1: Accuracy of different algorithms

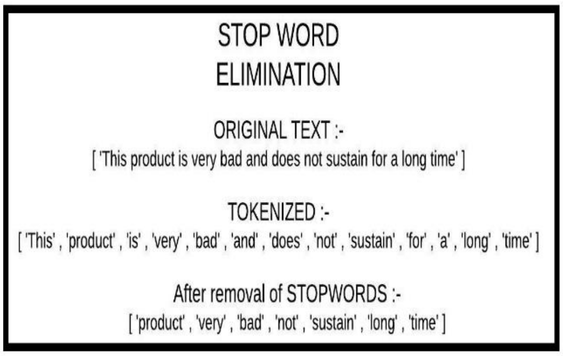
Firstly, the most important aspect of the project is the data set. So, the dataset is extracted by surfing over the internet. The dataset which best suits our interest is then selected. This dataset is then imported on the Machine Learning platform. This data is then distributed as training data as the entire data will be used. This data is the pre-processed using data cleaning methods like stemming and lemmatization word, stop word elimination, POS tagging, Word2Vec etc. With this cleaned data then we extract the various aspects and find the polarity of the aspects, then based on some predefined aspect category we start adding the various aspects to the category. We train the classifier and then the algorithm is applied to give an overall score for the aspects. Naïve Bayes’ algorithm is used for the opinion mining to achieve highest possible accuracy when aspects are considered.

A. Dataset

This is a list of over 5,000 consumer reviews for restaurant services like the service, Food, ambience and more provided by Restaurant from SemEval 2014 Dataset. The dataset includes columns like the reviews and another column with review classified as positive or negative. Along with this the project uses another dataset of 1,000 reviews taken from Kaggle, which contains reviews and their classification as positive, negative which is used for building a comparative model using Naïve Bayes Algorithm.

B. Data Pre-processing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data are often incomplete, inconsistent and / or lacking in certain behaviors or trends, and are likely to include many errors. Data preprocessing ways include Stop word elimination, removal repeated characters, hash tags and usernames, punctuations, lemmatization and also identifying the category in which the review falls. Also, another step in data preprocessing involves giving semantic meaning to a particular word which can be achieved using count vectorizer.



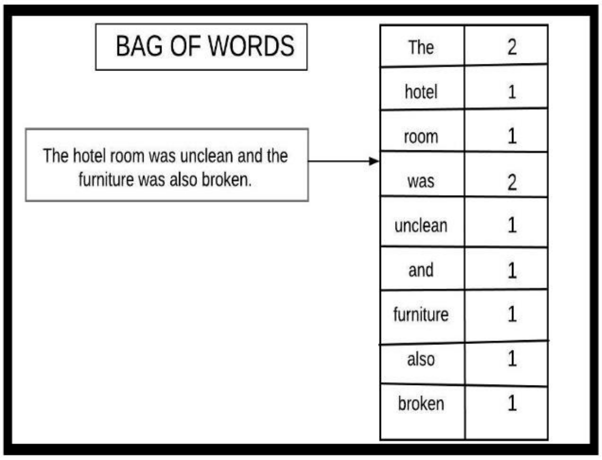
STOP WORD ELIMINATION

ORIGINAL TEXT :-
['This product is very bad and does not sustain for a long time']

TOKENIZED :-
['This' , 'product' , 'is' , 'very' , 'bad' , 'and' , 'does' , 'not' , 'sustain' , 'for' , 'a' , 'long' , 'time']

After removal of STOPWORDS :-
['product' , 'very' , 'bad' , 'not' , 'sustain' , 'long' , 'time']

Figure 2: Stop word elimination

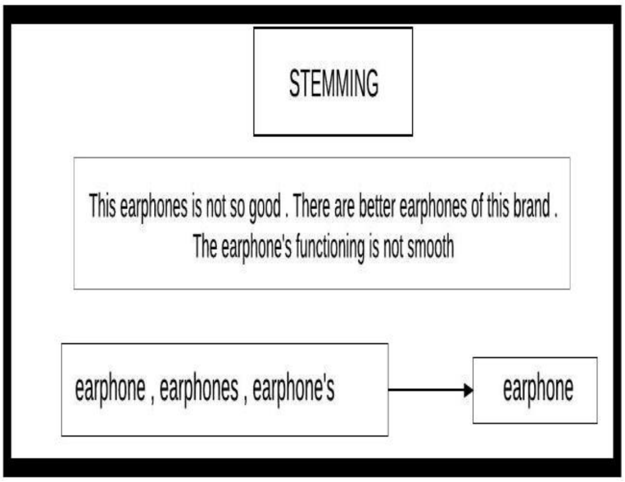


BAG OF WORDS

The hotel room was unclean and the furniture was also broken.

The	2
hotel	1
room	1
was	2
unclean	1
and	1
furniture	1
also	1
broken	1

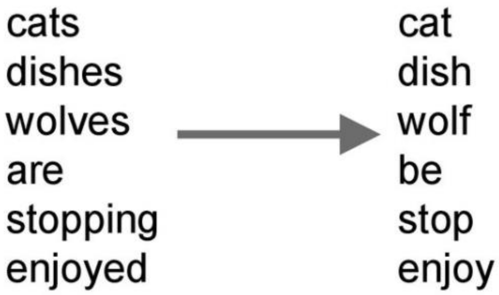
Figure 3: Bag of Words



STEMMING

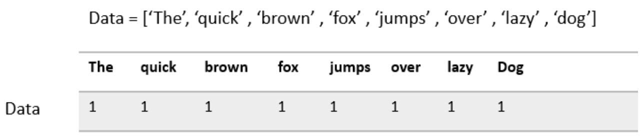
This earphones is not so good . There are better earphones of this brand . The earphone's functioning is not smooth

Figure 4: Stemming



cats	cat
dishes	dish
wolves	wolf
are	be
stopping	stop
enjoyed	enjoy

Figure 5: Lemmatization



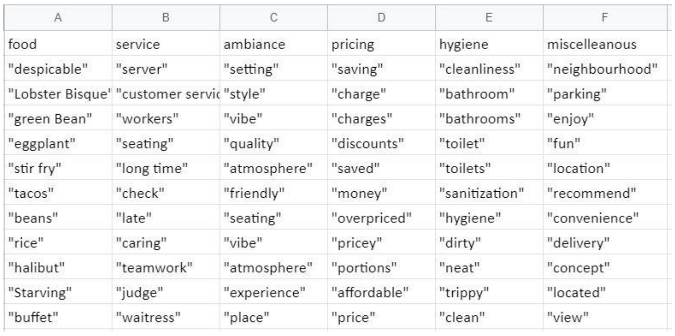
Data = ['The', 'quick' , 'brown' , 'fox' , 'jumps' , 'over' , 'lazy' , 'dog']

	The	quick	brown	fox	jumps	over	lazy	Dog
Data	1	1	1	1	1	1	1	1

Figure 6: Count Vectorizer

C. Aspect Identication

Assuming the reviews are grammatically correct, on the basis of topic modelling aspects can be extracted. A dictionary based approached is used for finding out the aspects which are present in a particular review. According to the six categories i.e ambience, food, price, service, hygience and miscellaneous a list of words belonging to each each particular aspect category are identified. Finally while parsing through the review any review which will be falling under the aspect list will be classified under that root aspect category.



A	B	C	D	E	F
food	service	ambiance	pricing	hygiene	miscelleanous
"despicable"	"server"	"setting"	"saving"	"cleanliness"	"neighbourhood"
"Lobster Bisque'	"customer servic	"style"	"charge"	"bathroom"	"parking"
"green Bean"	"workers"	"vibe"	"charges"	"bathrooms"	"enjoy"
"eggplant"	"seating"	"quality"	"discounts"	"toilet"	"fun"
"stir fry"	"long time"	"atmosphere"	"saved"	"toilets"	"location"
"tacos"	"check"	"friendly"	"money"	"sanitization"	"recommend"
"beans"	"late"	"seating"	"overpriced"	"hygiene"	"convenience"
"rice"	"caring"	"vibe"	"pricey"	"dirty"	"delivery"
"halibut"	"teamwork"	"atmosphere"	"portions"	"neat"	"concept"
"Starving"	"judge"	"experience"	"affordable"	"trippy"	"located"
"buffet"	"waitress"	"place"	"price"	"clean"	"view"

Figure 7: List of Words – for Aspects

D. Finding the polarities

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive or negative Natural language processing (NLP) is used for this purpose. It is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. To implement natural language processing in python Natural language Toolkit(nltk) was used. NLTK is one of the leading platforms for working with human language data and Python. In the following project we have used nltk to make predictions, that is, given a product review, a computer can predict if its positive or negative based on the text. Also, various packages are imported from nltk for the pre-processing step.

It starts by building a model for predicting the polarity of the reviews for this the data goes through various steps like stopwords removal, contraction removal, tokenization, lemmatization and stemming. Once the preprocessing is done the words are converted into vectors by using TF-IDF vectorizer. The reason for converting words to vectors is Naïve Bayes Algorithm can understand data in the form of vectors as if predicts on the basis of probability and each factor is independent of all the various other factors. In simple words the weight is assigned and semantic value is given to words on the basis of their overall occurrence. Once all this is done, a pipeline involving all these steps is created and the new reviews enter through this pipeline and then Naïve Bayes model predicts the polarity of reviews.

Example: I really liked the pizza in the restaurant but the service was disappointing.
Category: food -> pizza positive
Category: Service -> service negative
Once we get a score for every category, according to a threshold value set. The areas where improvements are required will be suggested. The overall working of our project can be explained through the following abstract diagram.

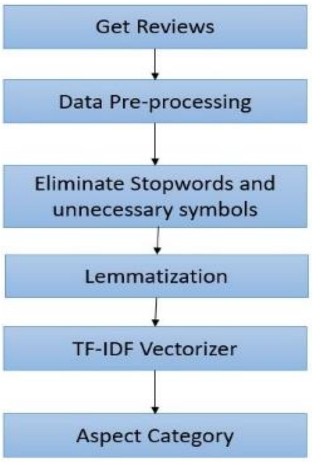


Figure 7: Data Preprocessing

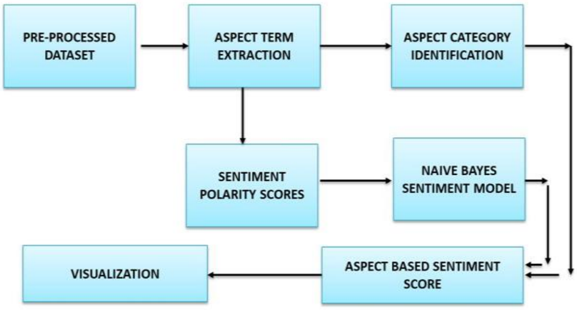

PRE-PROCESSED DATASET
ASPECT TERM EXTRACTION
ASPECT CATEGORY IDENTIFICATION
SENTIMENT POLARITY SCORES
NAIVE BAYES SENTIMENT MODEL
VISUALIZATION
ASPECT BASED SENTIMENT SCORE

Figure 8: Aspect Based Opinion Mining

5. IMPLEMENTATION

The implementation starts with cleaning the dataset by removing the unnecessary symbols, stopwords, punctuation and also converting words to their root words for ease to train the model and perform mining on it. Count Vectorizer along with TF-IDF are applied on the dataset so that sematic values of words can be found out. This preprocessed dataset is fed to Naïve Bayes model, which uses Gaussian Naïve Bayes following normal distribution. The result gives the accuracy of the model which is used for telling the sentiment for each review.

A list of related words for each of the five aspects namely ambience, food, service, price and hygiene are taken and then each individual review is parsed and by performing operations on the review and comparing it with the predefined list of words, each reviews aspect are found out that is what exactly the review is talking about. Along with this an additional aspect category called miscellaneous is taken in which all the reviews which don’t fall in the predefined category are put into miscellaneous.

Once all the aspects are found out, the opinion mining is done by collective aggregation of all the aspects and the most talked aspects are found out from all the reviews. Furthermore, using Vader library called sentiment intensity analyzer, is used which calculates the sentiment score for each review and also the score is normalized so that it can be easier to scale and evaluate reviews with respect to each other. While scoring various things are being considered like polarity, intensity, specificity.

Topic modelling is performed using Latent Dirichlet Allocation which finds out the five topics and allocates scores for each of the five topics in each and every review. LDA basically works by finding the topics inside document and then by finding words inside those topics.

Once all of this is done, the data is shown graphically for the various aspects which shows the number of reviews under each aspect. The collective scores for each aspect, based on the reviews are taken and areas for improvement if needed will be mentioned if the score is below a particular threshold value.

6. FEASIBILITY STUDY

A. Technical Feasibility

Programming Languages: Python
IDE: Eclipse, PyCharm
Frameworks: Flask, Bootstrap
Techniques: Data Mining, Machine Learning, Natural Language Processing
Web Technologies: HTML, CSS, JavaScript

B. Operational Feasibility

This System will be very easy to operate as the front-end would be very user-friendly as thus operationally feasible.

C. Legal Feasibility

The technologies being used to create this project are all open source and legally available. All the data being used will be taken from review portals which allow us to do so or from an establishment itself. The final product of this project does not infringe upon any rights. Any data will not be misused. Hence, this project is legally feasible.

D. Economic Feasibility

The cost of development of this project is very negligible. All the technologies being used to develop this project are open source, hence not costing us. All the data being collected online is also available publicly or with certain establishments being free to download. The very small amount of cost being spent by us is outweighed by the benefits of this project. Hence, we can say the project is economically viable for us to create.

E. Market Feasibilty

- This project will allow companies, hotels to reach large number of target groups electronically and simplify their process of market analysis.
- As it will be free of cost and user friendly, a small to very large set of opinions can be easily analyzed, so it can be useful for startups as well as reputed multinational companies • This project will be able to analyze user feedbacks and plan accordingly for the further development of product.

F. Schedule Feasibility

The project is divided into 4 phases, and each phase is assigned a time duration, along with buffer time so that the project gets completed, satisfying all the requirements and is ready to be deployed before the actual deadline.

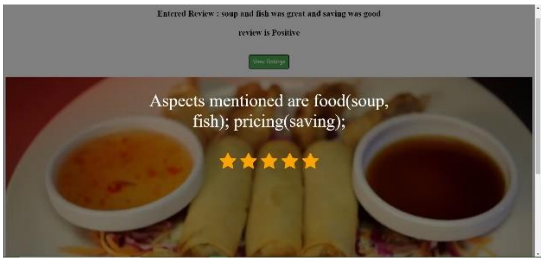
7. ADVANTAGES

The algorithm could also be used at any kind of platform with any kind of review as well if it works efficiently and with a lot of accuracy. It could also be used by any customer before visiting a restaurant to check whether the restaurant is good or not. The given opinion mining project identifies the aspects present in the particular review, gives the overall score for the particular review and also shows up a collective aggregation in the form of graphical representation of all the reviews on which mining has been performed and additionally it will suggest areas for improvement after the mining results. It can also be used by the restaurant managers and business owners to check upon the areas where improvement is needed.

8. RESULTS AND DISCUSSION

Thus, the restaurant reviews are successfully mined based on various aspects and their polarity, orientation and semantic information using Natural Language Toolkit, Latent Dirichlet Allocation and WordCloud. This gives the owner of the restaurant a clearer idea as to which areas should be improved and which areas should be kept consistent. The paper can be used at a greater level in the industry by using boosting LDA and RNN of various layers.

Evaluation is done by measuring accuracy, precision and recall. The total accuracy of 87.06% was obtained on the overall view of the dataset using Naïve Bayes Algorithm on manually annotated test dataset.



Entered Review : soup and fish was great and saving was good

review is Positive

Aspects mentioned are food(soup, fish); pricing(saving);

Figure 9.1: Review and Aspects Output

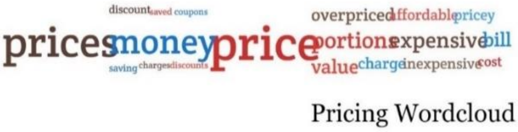


Figure 9.2: Analysis – WordCloud

9. FUTURE SCOPES

The paper uses analysed historical data and tries to predict the outcome and opinion mines them. The major benefit of the paper is that the result will be shown on the spot. The main aim of the project will be to increase the accuracy of the result so as to enhance the performance as compared to the previous approaches.

One of the advancements that can be done in this paper is that to include the same and also the analysis of different type of sentences like conditional, comparative sentences in order to improve the accuracy of opinion mining and also include automatic grouping of aspect synonyms. In this work the priority scores of opinion words are assigned using SentiWordNet, future work aims to avoid the SentiWordNet score, and find an unsupervised approach.

The paper has a number of applications. The major benefit of the paper is that the result of classification will be shown instantly. The paper can be used on any platform where in the originality of the reviews is to be known. Also, another improvement that can be made is that previously we considered all the grammar to be entirely correct but in real world this is not possible. So, identifying the proper grammar can also be done.

10. CONCLUSION

Opinion mining or sentiment analysis is a comprised area of natural language processing, computational linguistics and data mining, in it reviews about a topic is analyzed and expressed opinions are extracted. Most of the previous work is in the field of document or sentence level analysis. This paper proposes a new different syntactic approach to aspect level opinion mining, which use aspect dictionary, SentiWordNet, Dependency parsing, adverb adjective, adverb verb combinations, adjectives and adverbs together for opinion mining process with automatic acquisition of aspects. It is a syntactic based approach hence there is no need any training data. In the proposed method aspect matched opinion Words are extracted using dependency parsing. Polarity of opinions positivity or negativity of an aspect is found out using SentiWordNet, adjective, adverb adjective and adverb verb combinations using this aspect based visual summary can be produced, which shows positiveness and negativeness of each aspect from total reviews. The performance of the proposed method is evaluated by building an annotated test set of restaurant reviews. In this work only explicit aspects are considered and aspects are extracted using training. If modified the aspect extraction task without training will improve the accuracy. Word sense disambiguation is ignored in it.

11. REFERENCES

[1] V. Dhanalakshmi , Dhivya Bino , A. M. Saravanan "Opinion mining from student feedback data using supervised learning algorithms", MEC International Conference on Big Data & Smart City (ICBDSC) , vol. 3, April 2016, DOI: 10.1109/ICBDSC.2016.7460390, ISBN 978-1-5090-1365-4.

[2] Chinsha T C and Shibily Joseph ,"A syntactic approach for aspect based opinion mining", IEEE 9th International Conference on Semantic Computing , vol. 1, pp. 24-31, February 2015,

DOI:10.1109/ICOSC.2015.7050774 , ISBN- 978-1-4799-7935-6.

[3] Amiya Kumar Tripathy, Revathy Sundararajan , Chinmay Deshpande , Pankaj Mishra , Neha Natarajan "Opinion Mining from User Reviews" , International Conference on Technologies for Sustainable Development (ICTSD), vol. 9 , March 2015, DOI: 10.1109/ICTSD.2015.7095904, ISBN :978-1-4799-8187-8.

[4] Krishna B Vamshi ; Ajeet Kumar Pandey ; Kumar A. P. Siva "Topic Based Opinion Mining and Sentiment Analysis" , International Conference on Computer Communication and Informatics(ICCCI), vol. 8, Jan 2018, DOI :10.1109/ICCCI.2018.8441220, ISBN:978-1-5386-2238-4.

[5] Rushabh Shah and Bhumit Patel "Procedure of opinion mining and sentiment analysis" , International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 –5161 , Vol 4, No. 6 Dec 2014

[6] Alexander Pak and Patrick Paroubek "Twitter as a corpus for sentiment analysis and opinion mining.", Proceedings of the International Conference on Language Resources and Evaluation, vol. 7, pp. 1320-1326, May 2010.

[7] Jovelyn C. Cuizon, Jesserine Lopez and Danica Rose Jones "Text Mining Customer Reviews for Aspect Based Restaurant Rating", International Journal on Computer Science and Information Technology (IJCSIT), vol. 10, No. 6 Dec 2018.

[8] Shilpi Chawla, Gaurav Dubey and Ajay Rana "Product Opinion Mining using sentiment analysis for smartphone reviews", International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), vol. 6, April 23, 2018.

DOI: 10.1109/ICRITO.2017.8342455,

ISBN:978-1-5090-3013-2.