# Sentiment Analysis of Customer Text Content in Product Reviews Using the LDA Method

**Kishorchand Hirasingh Rawat[1], Dr. K.Amuthabala[2], Dr. Sasidhar Babu Suvanam[3], Prof. Keerthana P[4]**

*School of Computer Science and Engineering, REVA University, Bengaluru, India*

[1]*kishore.rawat@gmail.com*

[2]*amuthabala.p@reva.edu.in*

[3]*sasidharbabu.suvanam@reva.edu.in*

[4]*keerthana.P@reva.edu.in*

**Abstract**

In the current world of ubiquitous usage of Internet for most of our daily activities, purchasing of products and services from popular e-commerce websites is a well-known fact. Most of the consumers prefer to make an informed decision while making this purchase, for which they refer to the previous customer's reviews and feedback on the various features of the products or services offered. Hence it is imperative for every e-commerce organization to analyze this feedback to extract the appropriate sentiment of the customers, which then drives their recommendation systems as well as help in improving quality of the products and services. As customer reviews may contain user's mixed feedback on different features of the product, extracting sentiments becomes a challenging task. In this paper we propose aspect-oriented sentiment analysis using topic modeling algorithm called Latent Dirichlet Allocation (LDA). We extract the topic words from the customer reviews and map them onto various aspects or features of the product to perform aspect-specific sentiment analysis. The proposed approach combines word level and syntactic-relation level language analysis for an enhanced level of sentiment analysis. The results of this combined approach are much improved compared to conventional methods of sentiment analysis.

***Keywords***: sentiment analysis, topic modeling, Latent Dirichlet Allocation, LDA, review aspects, aspect-oriented sentiment analysis

## 1. INTRODUCTION

The usage of online e-commerce platforms, social media networks, online forums and blogs has increased tremendously in recent years, with increasing need for analysis of the feedback and reviews of the users of these platforms. Analysing these reviews and feedbacks in terms of positive or negative sentiments is critical to analyse the behaviour of the users, where these reviews express the customer's opinion about buying a certain product, services offered in a hotel, watching a movie etc. Sentiment analysis of these reviews becomes a very valuable resource for the business as well as future customers of the product or service.

Sentiment analysis consists of extracting the feelings and opinions of people from the review text, which is a huge challenge due to the scale and distinct structure of the language constructs. It involves usage of natural language processing, various statistical methods and optimized machine learning algorithms to extract the proper sentiments from the text corpus. There are various methods of sentiment analysis which work at document-level and sentence-level of the corpus. In this article, we propose to extract relevant aspects from the review

comments using topic modeling with LDA and then doing sentiment analysis on those topics using syntactic evaluation and polarity detection.

Aspects refer to the main topics or features of a product or service such as “engine” for a car review, “lens” for a camera, “food” for a restaurant and so on. Aspect based sentiment analysis provides much finer-grained opinion of users than the conventional coarse-grained document or sentence level analysis. The task complexity increases as one single user review can have multiple sentiments about each individual feature of the product, which requires the analysis of part-of-speech as well as any inherent syntactical meanings which signify the correct user sentiment. There are some known drawbacks of these methods like extracting large number of aspects some of which could be irrelevant to the domain of study. Also, extracting higher frequency aspects lose the infrequent aspects although some of them could be important for the analysis.

The following part of the paper is organized as follows: Section 2 provides details on the related work in this area, Section 3 illustrates the workings of the Latent Dirichlet Allocation (LDA) algorithm, Section 4 describes the approach and methodology of this paper, Section 5 illustrates the results and evaluation of the same. We conclude with Section 6 highlighting the important aspects of this paper as well as the interesting future work possible in this area.

## 2. RELATED WORK

In this section, we review some prominent work done on sentiment analysis specifically using the aspect-based topic modeling approach that is relevant to the study of this paper. The aspect-based sentiment analysis was first proposed by Hu and Liu [1] which gained wider popularity for further research in this domain. The proposal in [2] uses semi-supervised approach to use both labeled and unlabled data for the topic extraction process. The most popular method for extracting the aspects is the frequency-based method as illustrated in [3], which is used by many researchers despite being relatively simpler than other complex methods. This method works by extracting the high frequency words, most commonly which are nouns and noun phrases, and designate these as the candidate topics. After the aspect is extracted, the nearest adjective of this aspect is selected as its sentiment word. The work of Mubarak et al. [4] is among the recent works which uses this frequency-based approach. Akthar et al. [5] proposed the two-step approach for aspect-oriented sentiment analysis, viz. first extracting the aspect term and then in doing the sentiment classification in the second step. The rule-based method or syntactic relation-based method uses language specific syntactic structure and relations of words to identify the sentiment, which is the Double Propagation approach used in Qiu et al. [6].

## 3. LATENT DIRICHLET ALLOCATION (LDA)

LDA is a probabilistic generative model proposed by David Blei et al. [7], and it the most popular topic modeling algorithm used to extract topics from a large text corpus. LDA takes into consideration that each topic consists of a mixed set of words and each document is a mixture of a set of topic probabilities.

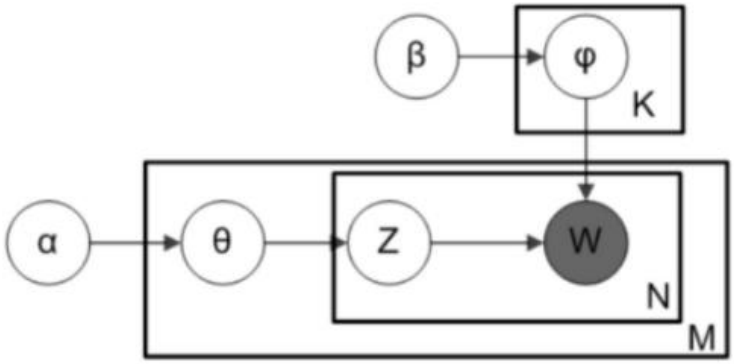


Fig.1: LDA Topic Modeling

We can understand the LDA generative process with the help of the different parameters, which are as follows,

- M – number of documents
- N – number of words in each document
- K – number of topics we want to extract from the corpus
- α – Dirichlet prior concentration parameter which specifies the document-topic density, higher alpha value would mean a greater number of topics per document
- β – similar to α for per-topic word distribution, higher beta would mean topics contain a large number of words
- z, w – multinomial distributions
- ϕ(k) – word distribution for topic k
- θ(i) – topic distribution for document i

## 4. METHODOLOGY

Customer review analysis allows companies to understand customers' needs and preferences and improve their services or create new products based accordingly. This can help businesses with better customer retention and acquisition, and hence revenue growth. Customer reviews are often multidimensional. For example, a restaurant customer could have a positive view on waiters and servers but might be less satisfied with the menu and the variety of choices offered by the restaurant. Aspect-based sentiment analysis allows to separately analyze each category, which corresponds to a specific component of the services/products. This method is also helpful in rapid sorting of customer complaints and assigning customer support tasks, which means effective customer support. The review analysis pipeline for this project included the following steps:

1. Extracting topics/aspects of a review
2. Sentiment analysis for each aspect
3. Extracting the subject of the review and its descriptors

**Dataset**: The dataset comes from restaurant visitors' reviews and included 3149 and 400 labelled reviews in the train set and test sets, respectively. Each review is labeled from a total of 8 aspects: 'food', 'menu', 'service', 'place', 'price', 'miscellaneous', 'staff', 'ambience'.

**Analysis**

1. Aspect identification:

There are two options for identifying aspects in a review:

1) Supervised learning using labelled dataset, and

2) Unsupervised topic modelling.

The first approach can be implemented using a multi-label classification algorithm, since a review can contain multiple aspects.

The second approach can be implemented using any of the typical topic modelling methods, such as Latent Dirichlet Allocation, non-negative matrix factorization, etc.

The extracted topics can then be cross-referenced with aspects, for example using the cross-correlations between topics and aspects, to identify the aspects associated with them. For this project, a voting ensemble was used to improve the prediction power.

The ensemble of the following 4 models was used to identify the main topic(s) for each review:

- multi-label classification (supervised)
- Latent Dirichlet Allocation
- non-negative matrix factorization with Frobenius norm
- non-negative matrix factorization with Kullback-Leibler divergence

1.1. Supervised aspect detection using multi-label classification: Classifier chain algorithm, with logistic regression as the base model, was trained on the labelled data (train set). The model achieved average scores of 0.86, 0.79, 0.82 for precision, recall, and f1-score, respectively, on the test set.

1.2. Topic modeling: For topic modelling, two methods were examined: Latent Dirichlet Allocation (LDA), and non-negative matrix factorization (NMF). LDA is a Bayesian method which finds topic using expectation maximization, where each topic can be represented by a group of words. NMF, on the other hand, is dimensionality reduction methods which decomposes the original the input data and transforms it obtain a smaller matrix, with fewer number of features.

For topic modelling, this method is used to find a few topics from the original feature vector (vectorized text). The decomposed matrices can be used to reconstruct the original matrix and the loss is calculated based on the reconstruction error and can be evaluated using Frobenius norm or Kullback-Leibler divergence.

2. Sentiment analysis: Sentiment analysis was performed using NLTK library (VADER) to estimate the positive, negative, or neutral sentiment associated with a review.

Extracting review subject and descriptors: This final step of the pipeline provides more contexts about the review through information extraction. To understand the review more specifically, it's useful to extract the subject of the review and its descriptors.

5 Results and Evaluation

The topic modeling results of the proposed approach is as show in Fig 2 below,

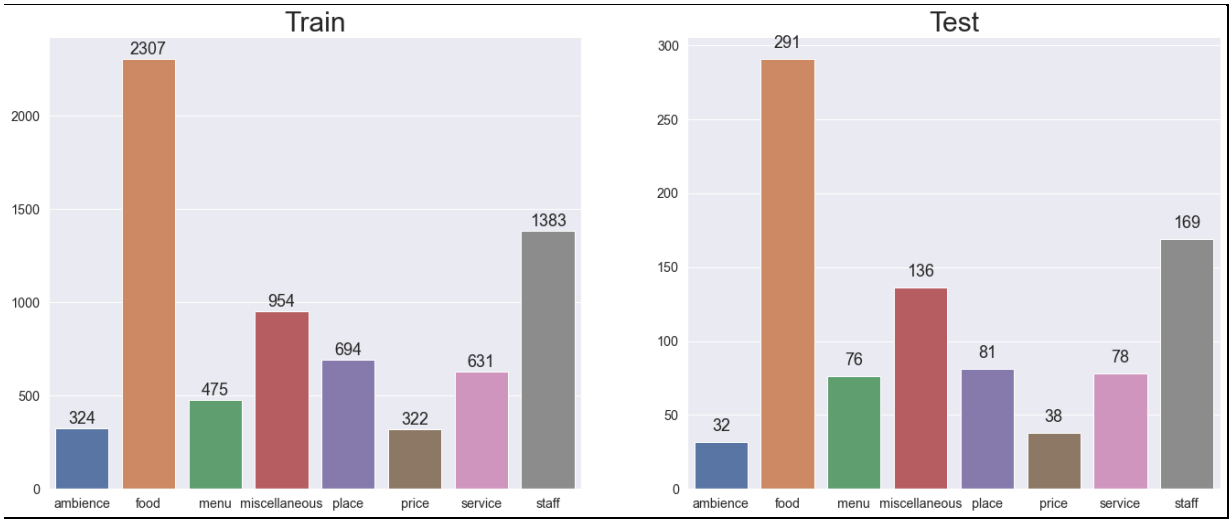


Fig.2: Topic Modeling of the Restaurant Reviews Dataset

The extraction of various topics from the review dataset is depicted using the word cloud image in Fig. 3,

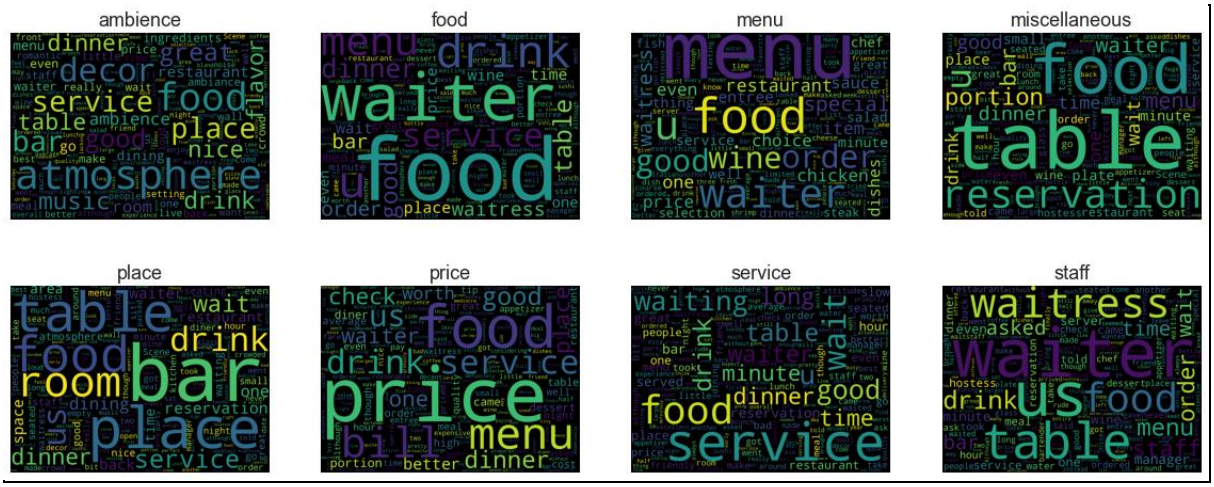


Fig. 3: Word Cloud for the Topic Extraction from Restaurant Reviews Dataset

The results in terms of precision, recall and f-measure show that our proposed method performs much better than the conventional baseline methods.

*Accuracy* = (*TP* + *TN*)/(*TP* + *TN* + *FP* + *FN*)

*Precision* = *TP*/(*TP* + *FP*)

*Recall* = *TP*/(*TP* + *FN*)

## 5. CONCLUSION AND FUTURE WORK

This paper has proposed an efficient approach for sentiment analysis of user reviews, showing the advantages of a fine-grained review analysis. This approach has used the LDA algorithm for topic modeling for the review corpus which extracts the most important “topics” from the dataset, and then maps these topics to the various aspects of the service. Then we perform sentiment analysis using various NLP and machine learning techniques. For the supervised aspects of the task we have to manually map the reviews to certain topics, which can be improved by automation in the future. The scope of scaling the algorithm to bigger size of corpus with improved performance using multi-core algorithms also could be a potential future work.


## ACKNOWLEDGEMENTS

Firstly, I wish to thank my guide **Dr. K. Amuthabala**, Associate Professor, School of CSE, for her constant suggestions for improvement during the reviews of the document and assisting in finalizing the document to its current structure through proper guidance and encouragement. I'd like to express my heartfelt gratitude to **Dr. P. Shyama Raju** Chancellor, REVA University for providing us congenial environment and surroundings to work on. I wish to convey on record my grateful thanks to **Dr. M. Dhanamjaya** Vice chancellor, REVA University, for providing excellent facilities in the campus, which helped us in every way during the preparation of this paper. I would also like to offer many thanks to the entire faculties who have encouraged me throughout the course of our Masters. I am also obliged to my Parents, family members and friends without whom I would not have been here. Their constant love and affection has given me the strength to complete this task.


## REFERENCES


[1] M. Hu and B. Liu, ``Mining and summarizing customer reviews,'' in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168_177.

[2] A. Hussain and E. Cambria, ``Semi-supervised learning for big social data analysis,'' Neurocomputing, vol. 275, pp. 1662_1673, Jan. 2018.

[3] Z. M. Zohreh Madhoushi, A. R. Hamdan, and S. Zainudin, ``Aspect-based sentiment analysis methods in recent years,'' Asia_Paci_c J. Inf. Technol. Multimedia, vol. 8, no. 1, pp. 79_96, Jun. 2019.

[4] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, ``Aspect-based sentiment analysis to review products using Naïve Bayes,'' in Proc. AIP Conf., 2017, Art. no. 020060.

[5] M. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, Feature selection and ensemble construction: a two-step method for aspect based sentiment analysis, Knowl. Based Syst. 125 (2017), 116–135.

[6] G. Qiu, B. Liu, J. Bu, and C. Chen, ``Opinion word expansion and target extraction through double propagation,'' Comput. Linguistics, vol. 37, no. 1, pp. 9_27, Mar. 2011.

[7] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003), 993–1022.

[8] Basha, Syed Muzamil, and Dharmendra Singh Rajput. "A supervised aspect level sentiment model to predict overall sentiment on tweeter documents." *International Journal of Metadata, Semantics and Ontologies* 13.1 (2018): 33-41.