Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 3001- 3007
Research Article

Opinion Mining With Hotel Review using Latent Dirichlet Allocation-Fuzzy C-Means Clustering (LDA-FCM)

S.N.Geethalakshmi[a], S.Shaambavi[b]

[a] Professor in Computer Science, Avinashilingam Institute for Home Science and higher Education for Women
[b] Enterpreneur Maison Grey
[a] sngeethalakshmi@gmail.com, [b] shaambavikct@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Indian tourism plays a significant role in Indian economic growth. The hotel booking and other stay-related inquiries are more important for every tourism manager. The tourist from the various country books their hotels based on the rating and reviews. The most remarkable improvement achieved through online hotel booking by customer review and rating. Customer opinions are more important for the growth of every business. This paper modelled the opinion mining system for hotel review. The wealth of customer thoughts and feelings has been identified by the review, and it also has a chance for the customer to reach their genuine opinion on social media. Most customers make online reviews and ratings. The opinion mining makes it very simple with a natural language process (NLP) that computers can easily understand every customer's human feelings and emotions. In this work, the NLP techniques of Part of Speech (POS) tagging, LEXICAL analysis, Latent Dirichlet allocation (LDA) has been used for an efficient opinion mining process. The fuzzy c-means clustering (FCM) has been used to cluster the opinion's positive and inclusive class determination. The proposed LDA-FCM based model works more efficiently than the conventional FCM algorithm. The performance has been evaluated by using Accuracy, precision, recall, and f1-score. The performance has been compared with the related work..

Keywords: Opinion mining, POS tagging, Lexical analysis, Latent Dirichlet allocation, fuzzy c-means clustering

1. Introduction

India is considered the seventh biggest country on the planet. India encountered multicultural evolution, and with a rich legacy and crowd attractions, the nation is among the most mainstream vacationer locations on the planet. Tourists from the various countries are visiting India countless every year [18, 14]. The possibility of staying may only depend on the level of hotels. The visitors may book hotels online hotel booking system based on ratings and reviews. The customer review and rating have been more important. It's become a normal and powerful piece of the booking order [17]. Therefore, it's likewise become a significant piece of the brand and advertising techniques. Power Reviews research finds that 95% of tourist counsel rating and surveys while booking hotels. What's more, 86% look at review as a fundamental asset when deciding between the hotels. Rating and reviews rank high among the top variables affecting booking choices, coming simply behind the cost because tourists trust them. The previews, customer posted opinions and conclusions posted online reliably positions among the best sources of data. The investigation shows that 66% of worldwide respondents trust shopper views posted on the web [2].

With the digitization and growth of the web from the last few decades, the opinion emanated for all like Twitter, Facebook, e-commerce sites, etc., an opinion makes dual benefits for both customer and marketer [16]. For the customer, the opinion makes efficient movement for every work like purchase, booking and decide about particular, and for marketer know about the customer feedback to improve the business [6]. But the huge number of opinions makes the job very critical. To mining, the review and rating the opinion mining and sentiment analysis take place to make a better decision about the large volume of review and rating for a single source [9, 11].

2. Literature review

There has been plenty of research carried out on opinion mining and sentiment analysis for various kind of reviews and ratings. In 2019 Guo et al. [10] proposed Tourist satisfaction analysis using Latent Dirichlet allocation and perceptual mapping. The work concentrated on the customer rating about the hotel stay. In 2020 Swagato et al. [5] deals with the opinion and emotional mining about online hotel booking. In 2020 Jia, Susan Sixue [13] proposed opinion mining for the tourist manager from china and the USA has been constructed with the restaurant customer review and rating the cross-cultural comparison has been taking place with the latent Dirichlet allocation and frequency analysis. In 2020 Da'u, Aminu et.al. [7] deals with opinion mining with the weighted customer about the product.

The rating has been considered with the review comments, and the word embedding POS tagging and many technologies of convolutional neural network CNN and collaborative filtering methods have been used in this work. All the work considered only the positive and negative opinion.

This paper concentrated on hotel reviews and opinions about Chennai hotel booking. The data has been discovered from the Kaggle repository. This dataset contains 4000 reviews collected from 500 hotels across Chennai. In this paper, many efficient methods have been used to make the opinion in Ming efficient. The method includes pre-processing with filtering, NLP, and the opinion has been clustered using enhanced fuzzy c-means clustering.

3. Proposed Mining Architecture: Lda-Fcm

Pre-processing

Pre-processing is vital in any data-mining measure as it straightforwardly impacts the achievement of any data mining project [1]. It reduces the complexity by removing the unwanted information from the original dataset. Most of the review dataset contains noisy data with outlier and duplication presents on the outsourced data. These make the degradation quality of the outcome. In this work, NLP methods were used to remove the noisy and unwanted data. For most language-related work like text summarization, word processing, etc., stop word removal is applied as the major pre-processing task, in this method; the unused or useless data has been eliminated using stop-word removal [8]. Mostly the, a, an in has considered the useless word meaning as commonly used. The next filtering was converting short mobile language to normal English language [12]. Most of the mobile-based reviews contain short mobile language words like gd, std, ok, etc.

The rule-based and stochastic tagging has been applied to many opinion mining processes. Broad advancement of the web has prompted the creation of an enormous measure of client-produced information. This information comprises numerous helpful data. Manual breaking down this information and grouping feelings in them is a debilitating assignment. Along these lines' opinion-mining technique is required. The opinion-mining approach utilizes regular language preparing where Part-of-Speech (POS) Tagging is a significant part. The presentation of any NLP framework relies upon the precision of a POS tagger. Two principles give that influence the precision of POS tagger are unknown words and uncertainty [15].

It comprises distinguishing and examining the design of words. Dictionary of a language denotes the assortment of words and expressions in a language. The lexical investigation is partitioning the whole lump of text into sections, sentences, and words. in this work two most widely used lexical normalization has been applied the first one is Stemming, it is a simple guideline-based cycle of stripping the additions ("ing," "ly," "es," "s" and so forth) from a word. The second included the Lemmatization, which is a coordinated and bit by bit methodology of getting the root type of the word. It utilizes jargon (word reference significance of words) and morphological investigation (word construction and sentence structure relations).

Input selection

Latent Dirichlet allocation (LDA)

In basic Latent semantic analysis, one of the generative statistical models has been the most widely used distributive model with the singular value decomposition. LDA is utilized for removing points from text that empower effective preparing, particularly for huge information analysis. Blei, Ng, and Jordan first proposed LDA in 2003[4]. This technique defeats past strategy in extricating point, named probabilistic inactive semantic ordering. This model processes each word in an archive as an example from a combination model, where the blend parts are arbitrary multinomial factors that can be spoken to as subjects. Notwithstanding, this gives no probabilistic model at the archive level. There have been a few explore utilizing LDA for separating points. LDA strategy was utilized to recognize the shrouded subjects to find conversation subjects in trick missions, and afterward, these are contrasted with the points recognized in veritable missions. Audits were examined to look at two contending items through their qualities and shortcomings and give sensible data that can help the board exercises in their organizations. LDA was additionally utilized to get knowledge of ride-hailing specialist organizations dependent on focuses talked about by clients about their assessment, experience, opinion, and objections through online media.

Clustering

Fuzzy c- means Clustering

Clustering is one of the main exercises of customer opinion mining. In which the similar data have been grouped to form a cluster. The number of the cluster formed depends on the valuable information. Clustering aims to maximize the similarity between inter classes by minimizing the cluster similarity. Clustering has been applied

in many applications like pattern recognition, machine vision, medical applications, etc., the wide variety of clustering algorithms present.

Fuzzy Clustering is an excellent unsupervised method for the examination of information and development models. Fuzzy Clustering is more normal than hard clustering. Articles on the limits between a few classes are not forced to completely have a place with one of the classes. Rather, they are doled out participation degrees somewhere between 0 and 1, demonstrating their fractional participation. Fuzzy C-Means (FCM) is a data clustering procedure wherein every information point has a place with a group to a few degrees determined by an enrolment grade. This method was initially presented by Jim Bezdek in 1981 [3] as an enhancement for prior grouping strategies. It gives a technique for how to clustering data focuses that populate some multidimensional space into a particular number of various groups. The favourable primary position of FCM clustering is that it permits continuous participations of information focuses to groups estimated as degrees in [0,1]. This gives the adaptability to communicate that information focuses can have a place with more than one cluster.

FCM technique is an unmistakable clustering method, has been used in a broad scope of designing and logical disciplines, for example, medication imaging, design location, information mining, and bioinformatics. Taking into account the reality, the first created FCM utilizes the squared-standard to decide the similitude among models and information focuses, and it performs well just on account of clustering spherical Clustering. Besides, a few calculations are created by various creators' dependent on the FCM with the point of grouping a more general dataset.

4. Performance metrics

The performance of a classification problem could be easily measured, where the output may be of two or more types of classes. A confusion matrix is a two dimensional table with "Actual" and "Predicted" and that both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)" and "False Negatives (FN)" as given below -

Description of the terms allied with performance metrics are as follows −

True Positives (TP) − represents both actual class and predicted class of data point is 1.

True Negatives (TN) − represents both actual class and predicted class of data point is 0.

False Positives (FP) − represents an actual class of data point is 0 and the predicted class of data point is 1.

False Negatives (FN) − represents an actual class of data point is 1 and the predicted class of data point is 0.

$$ACCURACY = \frac{TP+TN}{TP+FP+FN+TN} \quad \text{----------(1)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{----------(2)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{----------(3)}$$

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

5. Experimental Result And Analysis

This paper concentrated on hotel reviews and opinions about Chennai hotel booking. The data has been discovered from the Kaggle repository. This dataset contains 4000 reviews collected from 500 hotels across Chennai. The table1 resents the sample review with the opinion tag.

Table 1: Sample Sentence with Opinion Tag

S. No.	Example Sentence	Feature	Modifier	Opinion
S1	The hotel stay is excellent.	Hotel stay	-	Excellent

S2	During the stay, the food quality was very poor.	Food quality	Very	Poor
S3	The hotel room and atmosphere very clean	Room Atmosphere	Very	Clean
S4	The food quality really nice, amazing, and awesome.	Food quality	Really	nice, amazing, awesome

In this Table 1, the feature and modifier have been extracted using pre processing techniques of POS-tagging and lexical analysis for the particular review. Here the unwanted word has been removed by stop-word removal. The modifier and the opinion make the review very clear and quality like sentence 4 S4, the reviewer opinion has described the food quality with extracted opinion word of nice amazing and awesome has been modified to “Really nice," “really amazing, "and “really awesome."

Table 2: Comparison on Opinion determination of the training dataset

Classifier Methods	Positive Class			Negative Class		
	Precision	Recall	F-score	Precision	Recall	F-score
NB	93	73.2	81.9	86.6	50.7	64
SVM	95.1	99	97	95.1	83.6	89
ME	94.6	98.6	96.6	89.8	81.4	85.4
k-NN	95.2	99.3	97.2	94.4	83.6	88.6
LDA-FCM	**96.1**	**99.4**	**98.1**	**95.3**	**84.1**	**90.1**

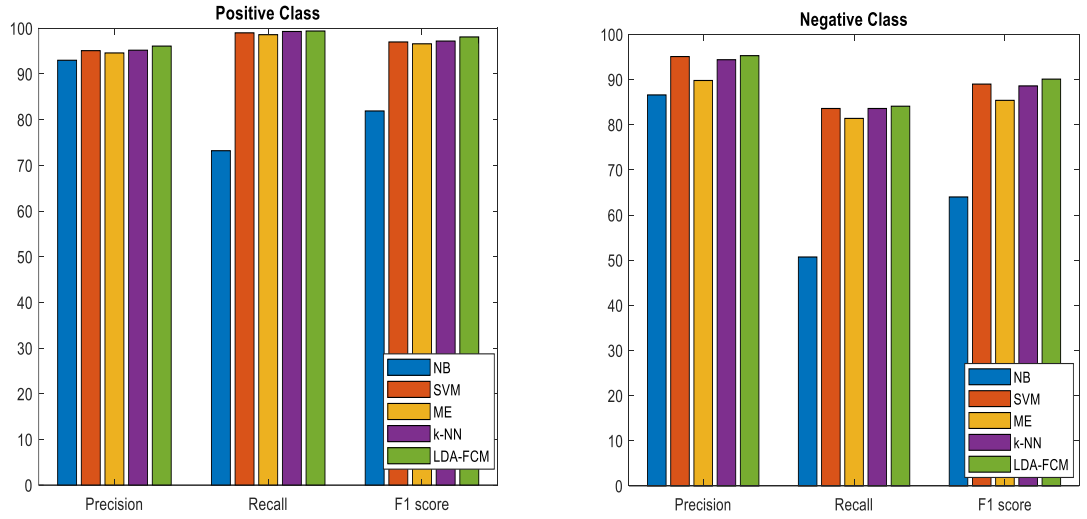

Figure 1: Positive and Negative opinion determination

Table 3: Comparison: Opinion determination of the testing dataset

Classifier	Positive Class	Negative Class

	Precision (%)	Recall (%)	F-score (%)	Precision (%)	Recall (%)	F- score (%)
NB	87.3	66	75.2	87.5	42.4	57.1
SVM	85.3	98.2	91.6	95.7	66.7	78.6
ME	85	96.8	90.5	88.5	69.7	78
kNN	85.2	97.9	91.1	91.7	66.7	77.2
LDA_FCM	**88.2**	**98.9**	**91.7**	**96.1**	**71.1**	**88.1**

The proposed model's performance has been evaluated with four different classifier model Naive Bayes (NB), Support vector machine (SVM), Max Entropy(ME), and K-nearest neighbour K-NN. Table 2 presents the comparison of 4 classifier models with the positive and negative class of review for the training dataset instance of 2800 from the 4000 reviews. Here the proposed model achieves high-performance metrics of precision, recall, and f-score for both positive and negative classes. The KNN has the second efficient method with the positive class, and SVM achieves the second efficient method on the negative class for training dataset review.

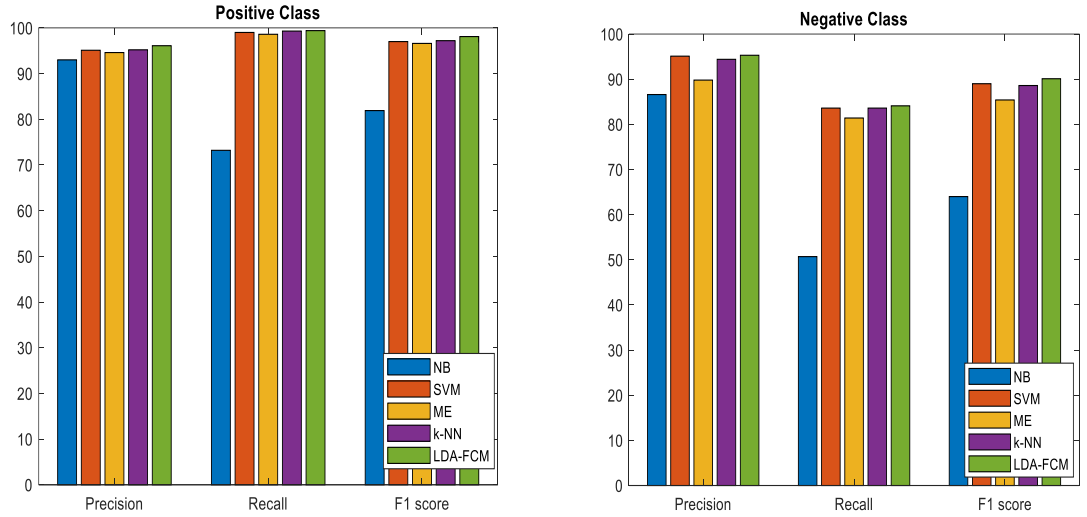

Figure 2: Performance analysis - Opinion determination positive and negative class

In Table 3 resents the comparison of 4 classifier models with the positive and negative class of review for the testing dataset instance of 1200 from the 4000 reviews. The proposed model achieves high-performance metrics of precision, recall, and f-score for positive and negative classes. The SVM has the second efficient method with the negative class for testing dataset review.

Table 4: Comparison results of precision, recall, and f-score values with other related work

Performance metrics	**Fuzzy Clustering**	**Fuzzy -C-means clustering**	**LDA-FCM**
Accuracy	96%	98%	**99%**
Precision	93%	95%	**96%**
Recall	88%	88%	**90%**
F1-Score	84%	85%	**88%**

The proposed LDA-FCA has been evaluated with the conventional method of the fuzzy and fuzzy c-means clustering method. The result shows that the proposed method outperforms the performance metrics of Accuracy 99%, precision 96%, Recall 90%, and f1-score 88%. Figure 3 describes the related work comparison.

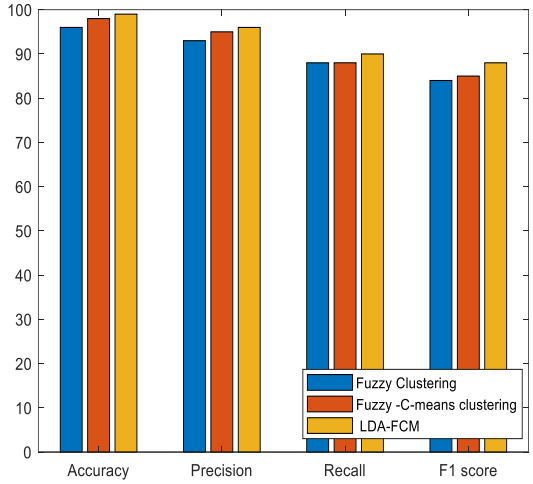

Figure 3: Performance Analysis with Related Work

6. Conclusion

This paper concentrated on hotel reviews and opinions about Chennai hotel booking. The data has been discovered from the Kaggle repository. This dataset contains 4000 reviews collected from 500 hotels across Chennai. In this paper, various natural language processes with efficient LDA methods and fuzzy C-means Clustering (LDA-FCM) have been used to achieve better performance. The 2800 instance of customer review used for training and 1200 has been used for testing purposes. The proposed LDA-FCA method outperforms the other related work outperforms. The performance metrics of Accuracy 99%, precision 96%, Recall 90%, and f1-score 88% have been noted from the proposed method..

References

1. Alnawas, Anwar, and Nursal ARICI. "Effect of word embedding variable parameters on Arabic sentiment analysis performance." arXiv preprint arXiv:2101. 02906, 2021.
2. Benlahbib, Abdessamad, and El Habib Nfaoui. "MTVRep: A movie and TV show reputation system based on fine-grained sentiment and semantic analysis." International Journal of Electrical & Computer Engineering, , 11, no. 2, 2088-8708, 2021.
3. Bezdek, James C. "Fuzzy c-means cluster analysis." Scholarpedia 6(7):, 2057, 2011.
4. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of Machine Learning Research 3, pp: 993-1022, 2003.
5. Chatterjee, Swagato. "Drivers of the helpfulness of online hotel reviews: A sentiment and emotion mining approach." International Journal of Hospitality Management 85, 102356, 2020.
6. Chen, Wen-Kuo, Dalianus Riantama, and Long-Sheng Chen. "Using a Text Mining Approach to Hear Voices of Customers from Social Media toward the Fast-Food Restaurant Industry." Sustainability 13, no. 1, 268, 2021.
7. Da'u, Aminu, Naomie Salim, Idris Rabiu, and Akram Osman. "Weighted aspect-based opinion mining using deep learning for recommender system." Expert Systems with Applications 140,pp: 112871, 2020.
8. Geler, Zoltan, Miloš Savić, Brankica Bratić, Vladimir Kurbalija, Mirjana Ivanović, and Weihui Dai. "Sentiment prediction based on analysis of customers assessments in food serving businesses." Connection Science, pp: 1-19, 2021.
9. Gour, Alekh, Shikha Aggarwal, and Mehmet Erdem. "Reading between the lines: analyzing online reviews by using a multi-method Web-analytics approach." International Journal of Contemporary Hospitality Management, 2021.
10. Guo, Yue, Stuart J. Barnes, and Qiong Jia. "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation." Tourism Management 59,pp: 467-483, 2017.
11. Ikasari, Diana, and Widiastuti Widiastuti. "Sentiment Analysis Review Novel “Goodreads” Berbahasa Indonesia Menggunakan Naïve Bayes Classifier." In Semnas Ristek (Seminar Nasional Riset dan Inovasi Teknologi), vol. 5, no. 1. 2021.
12. Jaitly, Gaurika, and Manoj Kapil. "A Productive Review on Sentimental Analysis for High Classification Rates." Progress in Advanced Computing and Intelligent Engineering, pp: 282-294, 2021.
13. Jia, Susan Sixue. "Motivation and satisfaction of Chinese and US tourists in restaurants: A cross-cultural text mining of online reviews." Tourism Management 78: 104071, 2020.

14. Johnson, David. "The Future Prospects of Tourism Industry: An Inquiry." International Journal of Humanities and Social Sciences Review (IJHSSR) 1, no. 1, pp.53-57, 2021.
15. Marzijarani, Shabnam Bagheri, and Hedieh Sajedi. "Opinion mining with reviews summarization based on clustering." International Journal of Information Technology 12, no. 4, pp: 1299-1310, 2020.
16. Möhring, Michael, Barbara Keller, Rainer Schmidt, Matthias Gutmann, and Scott Dacko. "HOTFRED: A Flexible Hotel Fake Review Detection System." In Information and Communication Technologies in Tourism, pp. 308-314. Springer, 2021.
17. Sheikh, A. A., T. Arif, M. B. Malik, and S. I. Bhat. "Extraction and Summarization of Reviews using Lexicon based Approach." In IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012-117. IOP Publishing, 2021.
18. Singh, Gurdeep, Vinay Duggal, and Taksh Gulati. Study of Tourism and Online Booking System of Hotels and Guides. No. 4866. EasyChair, 2021.