AI Assisted Sentiment Analysis of Hotel Reviews Using Topic Modeling

Sushma Ghogale
MSc Artificial Intelligence
Queen Mary University of London,
London,United Kingdom
Email: sush.ssg@gmail.com

Abstract— With surge in user generated content or feedback or reviews on the internet, it has become possible and important to know consumer's opinions about products and services. This data is important for both potential Customers and Business providing the services**.** Data from Social media is attracting significant attention and has become most prominent channel of expressing unregulated opinion. Prospective customers look for reviews from experienced customers, before deciding to buy product or service. Several websites provide platform for users to post their feedback for the provider and potential customers. However, biggest challenge in analysis such data is in extracting latent features and providing term-level analysis of the data. This paper proposes an approach to use topic modeling to classify the reviews into topics and conduct sentiment analysis to mine the opinions. This approach can analyse and classify latent topics mentioned by reviewer in on business sites or review sites or social media using topic modeling to identify importance of each topic. It is followed by sentiment analysis to assess the satisfaction level of each topic. This approach provides classification of Hotel reviews using multiple Machine learning techniques and comparing different classifiers to mine the opinions of user reviews through sentiment analysis. This experiment concludes that, Multinomial Naïve Bayes classifier produces higher accuracy than other classifiers.

Keywords—Latent Dirichlet Allocation (LDA) topic modeling, text classification, Sentiment Analysis

I. INTRODUCTION

Web is witnessing tremendous increase in amount of customer reviews in recent times. Review sites are able to get more reviews, ratings and opinions than the businesses themselves directly. Businesses are realising the importance of mining this data, as this data can make or break the business as more and more prospective customers are judging the product and services based on the past review comments or ratings [36][37]. Most sites even provide filter capability to exclude vendors or businesses with ratings below a threshold. With increase in magnitude of the reviews, complexity and time required to analyse and interpret the data, it has become increasingly challenging for Businesses and prospective Customers. Technology is continuously being relied upon to recognise, classify, interpret and visualise the reviewer's data into understandable and actionable information. As it is difficult to go through segregate topics from large amount of data manually, therefore, there is a compelling need for an automated algorithm that can process large data set from documents and identify topics automatically [19].

Nowadays, NLP and Machine learning approaches are playing a vital role in predicting customer sentiments through tone of the text data. Their use in data analytics is increasing rapidly to get meaningful results and recommendations that give clear picture about the user experience. Machine Learning is also helping potential Customers to compare the prices, service etc. before purchase[41].

Through Machine learning, Hotel industry has been transforming tremendously by providing information about the quality of their services and continuously improvising facilities for travellers [43][42].

One of the most important way of classifying reviews or feedbacks are based on features. In this paper we will be using Latent Dirichlet Allocation (LDA) from Genism package along with Mallet (Machine Learning for Language Toolkit) implementation [25]. LDA is used most widely to extract topics by latent variables. Topics are represented as bags of words (Corpus), which represent specific features of corpora. Articles or sentences are allocated to appropriate topics by probability distribution of implicit weights of words[1].

Sentiment analysis is another important way of extracting features from user generated content (UGC). Sentiment analysis uses natural language process to analyse text considering specific situations. It helps to decipher user's feelings towards a topic, features or a thing.

User's feelings are categorised into positive, neutral and negative which correspond to like, mixed feelings, and dislike respectively. Sentiment analysis requires group of words to be analysed to identify sentiments, excitements and tone of the users for the product or services. This paper uses LDA to identify and extract topics and evaluation techniques to extract sentiment pairs. This paper uses LDA method to classify reviews followed by sentiment analysis on topics generated by LDA. LDA possesses the ability to handle large amount of data set by treating parameters as random variables [3][4].

II. LITERATURE REVIEW

A. LDA

Recent years have witnessed increased use of machine learning and natural language processing in LDA topic models. LDA is preferred over other topic modeling methods, which considers group of words in a document to have several topics. Probability of a topic is distributed among words in the subgroup [5].

LDA works by identifying pattern of words and assigning them to a relevant topic. Each topic has collection of group of words and each document is a group of topics which are distributed across the document. In this experiment, we are using LDA Mallet package, which is known for increasing the speed of the implementation of LDA [6].

In this experiment, we are using unsupervised topic modeling, which means LDA topics are not labeled. i.e. they don't have any prior content knowledge. Using unsupervised LDA technique means algorithm learns to identify topics from the data set and algorithm has to look for the keywords in the corpus to get the meaningful topic. This is different from supervised technique, where, the algorithm is provided with pre-labelled topics in the dataset [25].

B. *Sentiment Analysis*

Opinions are subjective to reviewer based on their experience. Opinions are usually hidden in sentences expressed by the reviewer. Challenge is to extract sentiment from expressions, which are not standardised and heavily depend on the language skills and interest of the reviewer. Sentiment analysis is a systematic and structured way of extracting subjective information like tone and sentiment especially in the form of positive or negative opinions about a topic [22]. Customer may provide positive, negative and neutral feedback/review comments. Sentiment analysis techniques are used to extract the excitement, anguish, mixed emotions kind of a sentiments from the feedback text provided [7][8]. Unsupervised learning algorithm LDA has been used in this experiment to categorise the text in the document into topics based on the user feedback. This is followed by sentiment analysis to assess the emotion through opinion mining [19]. In this experiment the emotions behind the reviews are categorized in to positive, negative or neutral depending on the reviewers' ratings.

C. *Proposed approach*

In this experiment, LDA was used for multilabel categorisation of the feedback reviews from the dataset, which in turn is used for sentiment analysis. Several researches have been done on using Bag of Words (BoW) for feature extraction. This has a weakness of long running time and results are not optimal because of accommodation of too many features [39]. In the proposed approach TF-IDF for feature extraction and text classification for sentiment analysis to improve accuracy are used. For LDA, Bag of words (BoW) technique is applied.

III. Related work

This research references extensive work done in the area of sentiment analysis, especially for the hotel industry. Guo, Y., Barnes, S. J., & Jia, Q. (2017) have researched that historically Customers' feedback analysis have been quantitative reviews of the response arising from ratings and qualitative customer feedback like questionnaire, which are cost effective.

However, these don't provide the visibility of dimensions of services which are important for the Customers. Their experience, excitement, anguish or any such sentiments against these dimensions are not captured or considered. Last few years have witnessed extensive study, research and technological development to capture and analyse large amount of User Generated Content (UGC) [50]. Alaei, A., Becken, S. and Stantic, B., 2017 have researched to show that the use of Bag of Words (BoW) approach is less effective for sentiment analysis. Research also shows that N-grams and complex linguistic features are beneficial when a large corpora of data is available for training the models [51]. Probability-based technique that assigns group of words to topics named Latent Dirichlet Allocation (LDA) was developed by Blei et al [5]. Girolami M and Kaban A 2003 proposed use of Latent Dirichlet Allocation (LDA) as generative approach to language modelling to overcome the inconsistent generative semantics of Probabilistic Latent Semantic Indexing (PLSI) [3]. Serra Cantallops, A. and Salvi, F., 2014 have proposed to use language processing (NLP) techniques like text summarisation and text classification to analyse reviews [52].

IV. Method

First step is to collect data and preprocess text. Second step is to perform LDA on cleaned data to get meaningful group of words as topics. Then topic coherence measure is used to select optimal number of topics. Because this is an unsupervised approach, model depends on coherence and perplexity, where lower perplexity would mean a good model fit and is generally considered the best model [25][26]. Next step is to find dominant topic for the set of reviews keywords through topics obtained from LDA. Finally, to get the reviewer's opinion, sentiment analysis is done on topics. The detailed discussion on the Method is discussed in following paragraphs. Figure 1 shows the workflow of the proposed approach of the model.

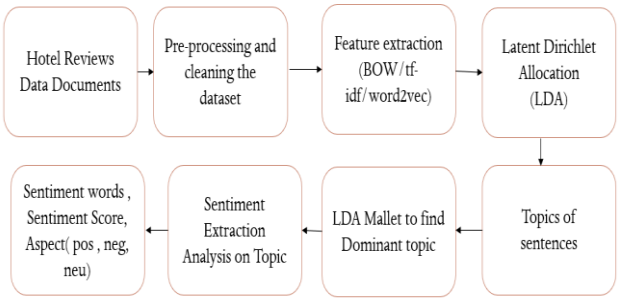


Figure 1 Proposed Approach

A. *Dataset and Preprocessing*

Dataset

Dataset used for this project is Hotel Reviews csv file which has more than 500K reviews from different Hotels around the world. The attributes we are using in this approach is mainly focused on ratings and reviews.

Preprocessing

At the preprocessing level, a mathematical representation of documents is made. That means the document is converted to machine understandable language (Binary), so that the classification of documents can be carried out by a machine. Hence Bag-of-Words (BoW) are formed [15]. Bag-of-Words (corpus) that is implemented in this project is information from 1st to nth token results, position of token words and position of token documents. The steps consist of tokenization, stopword removal, stemming, bigram, trigram and lammatisation.

1) Tokenisation

Tokenisation process is the first stage in pre-processing of an information retrieval. The process of tokenisation is to divide the rows of words from document into single words. The tokenisation technique also takeoff certain characters such as punctuation and changes all letters lowercase. For Eg, Location becomes location.

2) Stopword Removal

Few words in the document they appear frequently which has no meanings and in terms of Information Retrieval, these are called Stop-words [16]. Stop-words should be removed during pre-processing so as not to affect the indexing process.

3) Stemming and lammatisation

Stemming is the process of changing tokenised-words in a text document into the basic-level of words. Stemming is used in information retrieval process. Through Stemming, the effectiveness of information increases and also helps in reducing the size of the data. [15]. Where as in lammatisation words in third person are changed to first person and verbs in past and future tenses are changed into present tense.

4) Bigram and Trigram

Bigram is a sequence of two adjacent elements from a row of document of tokens and Trigram is a group of three consecutive elements from a row of document of tokens, that can be letters, words or syllables. We used Bigram and Trigram in this approach as they have an impact on the accuracy of the analysis, as extracting sentiment from a single word would reduce the intensity of the sentiment or even give an opposite meaning of the sentiment e.g. not happy, not bad, never been bad.

5) Bag of words (dictionary) and corpus

Prior to topic modeling we convert preprocessed text to a Bag of words(dictionary). In this dictionary, a word and its value are stored in the corpus, depending on the frequency of its appearance. [35][39].

B. LDA (Topic Modeling and Algorithm)

LDA with Gensim package, along with Mallet's implementation produces the group of words that belong to unique topics and runs faster resulting in better topic segregation. Topic modeling is process of creating list of words out of words from a dataset. A group of related words are clustered and are assigned a topic to identify the category or feature. This creates a mapping between a corpus of documents and group of clusters (topics). Now, let's say *n* is the number of topics on which reviewer is providing the feedback, though the value of *n* can be any number based on the sematic interpretation i.e for standard we choose *n=10.* Also the selection of number of words is dependent on the corpus, if we want to find uncommon (unique) topics from large corpus we select *n=100*. For this paper we choose *n=12* and each set of words in topic has weights, the words occur in topic in highest weighted word to lowest weighted word, that is in decreasing order. This helps in selection of keywords for topic. Topics are hand-classified or manually labeled based on the highest probability words [25]. Since LDA uses all text data from Hotel reviews for topic modeling, we are not splitting train and test data.

Topic labelling is done manually by reading the highest weighted words that belong to each topic. For example, if the word list in LDA is assigned as 'Station' 'near''location''walk''minute', we denote this as Location because these are group of words a reviewer might have used to describe the distance from points of interest. In the same way, we assigned other categories as Service, Food, Cleanliness, Staff, Facilities, Comfort, Price etc.

Figure 2 is the sample graphical representation of how the LDA picks the related words from a document and groups them into topics.

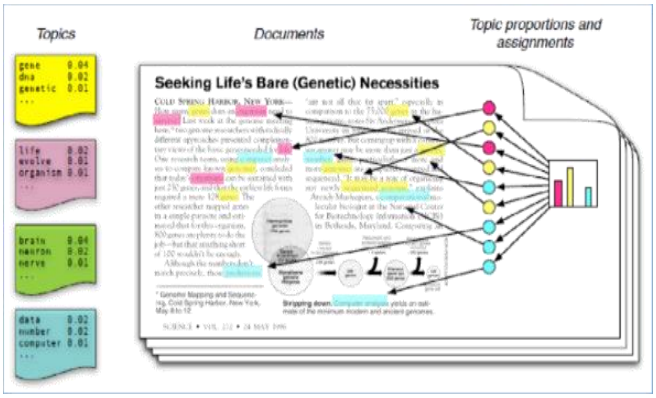


Figure 2 Example of how LDA works [20]

The fundamental principle of LDA topic level analysis is to generate group of words as a topic in an unsupervised technique.

Figure 3 shows the Graphical model of the LDA topic modelling approach of topic classification and generation. Where α and β are the hyper parameters belonging to corpus level which are taken one-time as sample data for analysis while generating Bag-of-words (corpus). The variable $\Theta(d)$ are the variables belonging to the document level, and samples are taken for analysis one-time in a particular document. The word level variables are represented as zdn and wdn, the samples are taken for analysis once for every word in the document. [17][18][56].

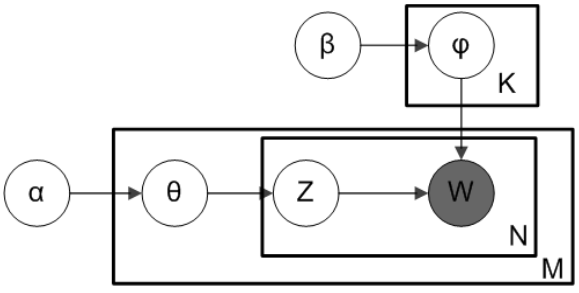


Figure 3 The graphical model of the LDA topic model.

In this unsupervised approach of finding the topics through LDA for Hotel Reviews, we use complete review data for

training purpose. This data is used to find Dominant topic by calculating topic coherence score.

2) Topic Coherence and LDA Dominant topic

Next step is to find the dominant topic from the topics generated from LDA. For this we use LDA mallet. LDA Mallet increases the coherence score and helps in finding the dominant topic in each review document sentence. From this we can find the topic number which has the highest percentage contribution in that document [14]. To get the number of topics which has high coherence score we use coherence measure.

3) Coherence Measure (UMass-coherence)

Coherence measure proposed in Natural Language processing is used to evaluate topic constructed by some topic model as LDA topic model. In other words, Coherence measure is discussed in scientific term as a formalism to quantify hanging and fitting together of information pieces [28][27][30]. It is defined as an average or median of pair of word similarities form top words of a given topic. Word similarity is grounded on external data which is not used during topic modeling [31]

UCI-coherence uses point wise mutual information (PMI) and word co-occurrence counts collected from Wikipedia based on a Boolean window mode. This word similarity measure is from bad to good topics that come closest to the quality of human judgements.

In an inferred posterior distribution, topics are visually analyzed on how well they fit in the real observations [6]. However, coherence measure is a proposed model to automatically judge interpretability of paired word sets. The coherence measure proposed by researchers [33] is also based on co-occurrences of word pairs. Given an ordered list of words T=(wl,..,wn) the UMass-coherence is defined as equation (1)

$$\mathrm{CUMass(T)} = \sum_{m=2}^{M} * \sum_{l=1}^{m-1} * \log \frac{p(wm,wl)+1/D}{p(wl)} \quad (1)$$

A Boolean model is probability of the ratio of number of documents of ‘wm and wl’ and the total number of documents in the corpus ‘D’. To avoid logarithm of zero value w ass smoothing count 1/D Coherence graph is shown in Figure 4.

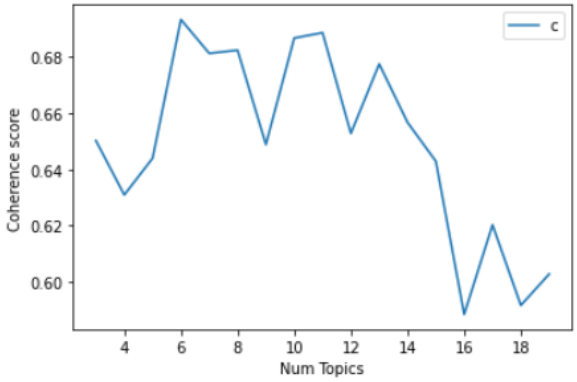


Figure 4 Coherence graph

Figure 4 shows different coherence values against the number of topics. The words within topics with high probability are words that tend to co-occur more frequently [21]. Figure 5 shows that the coherence score is highest, approximately more than 0.68, for topics between 5 to 7 and 10 to 11. Now for choosing optimal number of topics for LDA model with different values of number of topics (n) and pick the topic which gives highest coherence value we use LDA Mallet. Mallet chooses topics with high coherence value which usually are meaningful and interpretable topics [25][26]. In this proposed experiment the keywords (group of words belongs to topic) are being repeated in multiple topics. This means the choice of topic through Mallet is too large [23][24]. Table 5 shows the sample output generated topics (Dominant topics) with high coherence value and their respected labels.

Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Review_Title	Review Rating	dominant_topic_Label
0	10.0	0.2093	walk, close, station, easy, minute, train, cit...	pleasure this nights recently. This perfect ev...	5.0	Facilities,Location
1	11.0	0.3978	amazing, definitely, would, recommend, wonderf...	very lovely first visit this iconic bar! Wonde...	5.0	Service
2	5.0	0.3910	staff, location, breakfast, friendly, helpful,...	Rhodes Hotel nights, location taking Paddingto...	4.0	Staff
3	6.0	0.1827	feel, make, always, special, smile, home, welc.	Form moment arrived until left experienced abs...	5.0	Food, Service
4	0.0	0.6617	check, book, give, reception, night, ask, take.	Well strange London's 5star when comes along e...	1.0	Facilities

Table 1 Sample output of LDA with Dominant topic

V. Sentiment Analysis

Tourists look for affordable hotels with features and service levels important to them before booking. The trips might be family vacations or business trips. They look for reviews from travelers who have already experienced the stay [34]. As Sentiment analysis is a process of computationally identifying and categorising opinions and feedbacks expressed in the text, the opinion might be positive, neutral, or negative depending on the reviewer's experience. For aspect-based opinion mining we have used the dataset generated from LDA. The illustration of dataset is shown in the Table 1. As the Review ratings range between 0 to 5, we manually categorised ratings into three categories. Ratings which are rated above 4 are labelled Positive ratings which are rated below 2 are labelled as Negative and ratings between 2 and 4 are labelled as Neutral.

Figure 5 (a and b) shows the Graphical representation of reviews with respect to rating. The bar-graph shows that Positive reviews are comparatively higher than negative and Neutral reviews.

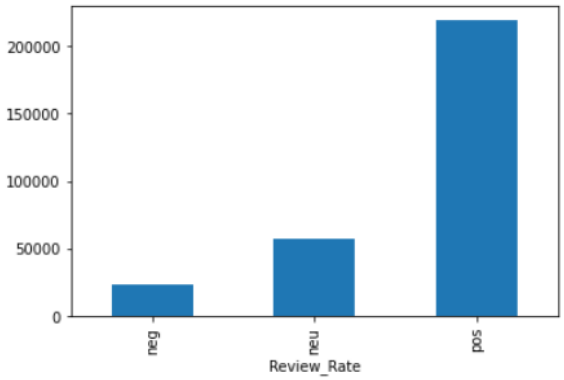


Figure 5 (a)

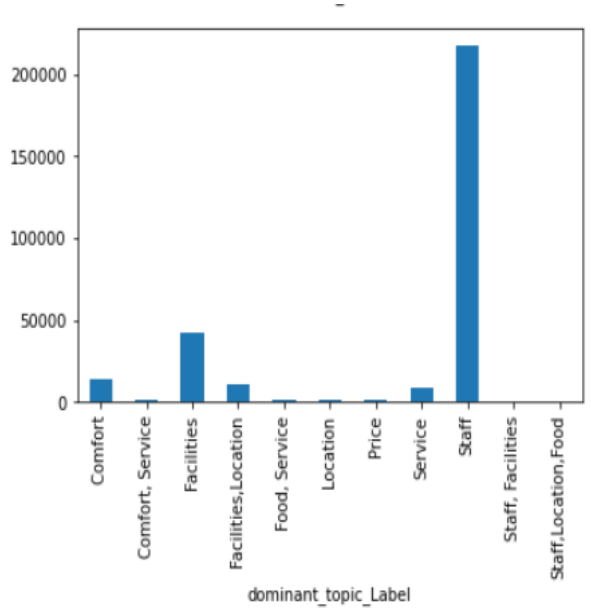


Figure 5 (b)
Graphical representation of review rating and Dominant topics

Next step is preprocessing the data in the same way as described in LDA approach. Additional TF-IDF is used instead of bag of words [41]. Finally, for classification of reviews we used Naïve Bayes, Support Vector Machine/Support vector Classification (SVM/SVC), Decision Tree and Linear Regression to find the best fit accuracy of the model for each topic (subgroups). Figure 6 shows the Machine learning approach for sentiment Analysis.

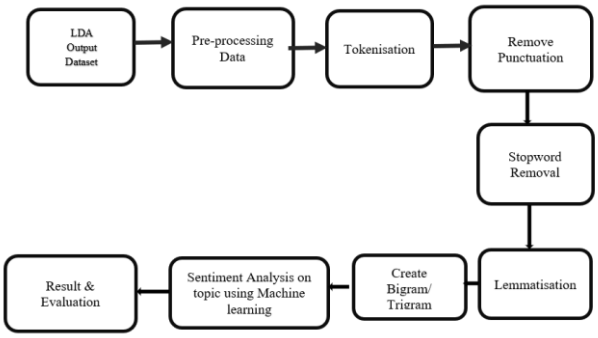


Figure 6 Sentiment Analysis approach

A. *TF-IDF Vecorisation*

TF-IDF stands for 'Term Frequency-Inverse Document Frequency TF-IDF creates a set of its own vocabulary from the data document. This technique is widely used when we try to extract some usual information from document or reviews data. TF-IDF assigns weight to the word based on how many times it occurred in the document set [41].

TF (Term Frequency): TF measures the frequency of a term (t) in a document. It is given by equation (1)

$$TF(t) = \frac{Number\ of\ times\ term\ t\ apperas\ in\ the\ document}{Total\ number\ of\ terms\ in\ the\ document} \quad (1)$$

Inverse Document Frequency (IDF): TF gives equal importance to all words but IDF measures how important a word is. IDF can be computed using equation (2) given below

$$IDF(t) = \frac{\text{Total number of documents}}{\text{Number of documents with word t in it}} \quad (2)$$

Finally, TF-IDF is obtained as the product of term frequency and inverse document frequency and is given by equation (3)

$$\text{TF-IDF} = \text{TF} * \text{IDF} \quad (3)$$

B. *Classification Methods*

1) Naïve Bayes Classification (NB): It is a simple probabilistic classification approach which is based on the Bayes probability theorem and naïve assumption of conditional independence [44][45]. In Naïve Bayes the conditional probability is calculated from training data set, provided the predictors are ordinal variables and we use Multinominal Naïve Bayes where the data represented as word vector counts.

2) Decision Tree Classification (DT): Decision tree classifier is a tree-shaped structure that represents sets of decisions. It starts with single node, that is root node, that pictures entire population. This root node usually splits into two sub nodes, represents terminal node or leaf node which does not split further. The construction of DT depends on the algorithm used [45].

3) Support vector machines (SVM): Support vector machine/support vector classification are algorithms that generates a set of hyperplanes to separates the data into multiple categories depending on the type of classification method [46][47][48]. SVM is usually used for linear classification.

4) Logistic Regression (LR): Linear Regression is the relationship between two or more variables. It minimises the residual sum of squares between the predicted and the target data by linear approximation [49]

C. *Evaluation*

Performance of multiple text classification and sentiment analysis is evaluated using accuracy, precision, recall, f1-score [53]. To calculate we use Tp (True Positive-Number of correct results that have been diagnosed correctly), Fn (False Negative-Number of incorrect results that have been diagnosed incorrectly), Fp (False Positive-Number of incorrect results that have been diagnosed correctly)and Tn (True Negative-Number of correct results that have been diagnosed incorrectly). In all criteria results range between 0 and 1, if the result near to 1 the model performance is accurate and gives better results [53][54].

Accuracy: It is the proportion of appropriate classified results to the entire population, shown in equation (4)

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (4)$$

Precision: Precision is the ratio of correctness of the results within the outputs. This indicates higher number of correct results and less incorrect results. shown in equation (5)

$$\text{Precision} = \frac{Tp}{Tp+Fp} \quad (5)$$

Recall: Recall is the ratio of correct results to the base of all results. This indicates high number of correct result and less number of incorrect results that has been identified incorrectly. Shown in equation (6)

$$\text{Recall} = \frac{Tp}{Tp+Fn} \quad (6)$$

F1-Score: This is the combination of Precision and Recall. Shown in equation (7)

$$\text{F1-score} = 2 * \frac{Precision\ x\ Recall}{Precision+Recall} \quad (7)$$

VI. Results and Discussions on Comparision techniques of classifiers of Sentiment Analysis

This section shows the detailed discussion of results from four different Machine learning types of classifiers for multi-class text classification on each topic (Sub-group). Experiments and comparisons conducted between Naïve Bayes, SVM, Logistic regression and Decision Tree classifiers, to classify sentiments into Positive, Negative and Neutral according to reviewer ratings on each topic (Sub-group).

The dataset is split into 80-20 for training and testing, respectively. In this approach we have evaluated the topics Service, Staff, Comfort, Price, Location, Facilities and Food by using aforementioned classifiers. Table 9 and 10 shows detailed comparison of results of the classifiers.

In this approach, we have built classification model and evaluated the predictions of training and testing data trying to identify which model performs well on each subgroup. Table 10 shows individual topic performance accuracy of each classifier and how it varies when trained on different subgroups.

Precision for three classifiers on topics show more than 90% because of high true positive reviews in this category. All positive samples are classified as positive samples and very few of the positive samples are classified incorrectly.

Recall for the topics staff, Location, Food, Service shows more than 70% and Recall for topics - Comfort, Price and Facilities show almost 60%. High value of Recall shows that the relevant reviews are successively retrieved. F1-score is also high for NB compared to other classifiers.

The constructed model accuracy varies with different classifiers. Naïve Bayes and Logistic regression classifier provide good results on the linear classification model.

In comparison, Support Vector Machine/classifier with linear model is very fast but it did not provide good classification accuracy in this experiment. NB performed at high precision, that is more than 90%, because true positive reviews were more than true negative reviews. On the other hand, precision and F1-score were very low for Decision tree classifier.

As we observe from the Table 2 and 3, Naïve Bayes and Logistic regression provided high accuracy accounting for more than 90% for the topic 'Food', while Decision Tree classifier generated least accuracy for topic 'Comfort' amounting to 40%. Overall, we conclude from experiment that Naïve Bayes is one of the best classifiers compared to other three classifier as it provides high accuracy in all subgroups(topics) compared to other three classifiers.

LR, NB and SVM performs well when the training data is less [54]. Decision tree is good for non-linear classifier and did not perform well in this experiment [54].

	Service	Staff	Location	Food	Comfort	Facilities	Price
Classifier							
Logistic Regression	**0.844**	**0.76**	**0.744**	**0.90**	**0.493**	**0.631**	**0.62**
Naïve Bayes	**0.844**	**0.76**	**0.742**	**0.90**	**0.525**	**0.636**	**0.63**
Support Vector Classifier	**0.833**	**0.75**	**0.713**	**0.90**	**0.448**	**0.608**	**0.55**
Decision Tree	**0.73**	**0.64**	**0.612**	**0.85**	**0.400**	**0.499**	**0.48**

Table 2 Comparison of Accuracy Results between various classifiers for sentiment Analysis on topic

Topic	Evaluation	Logistic Regression (LR)	NaiveBayes (NB)	Support Vector Machine (SVM)	Decision Tree (DT)	Dominant classifier
Staff	**Accuracy**	**0.76**	**0.76**	**0.75**	**0.64**	**NB & LR**
	Precision	**0.99**	**0.99**	**0.98**	**0.67**	
	Recall	**0.76**	**0.76**	**0.75**	**0.64**	
	F1_score	**0.86**	**0.86**	**0.85**	**0.66**	
Location	**Accuracy**	**0.74**	**0.74**	**0.71**	**0.61**	**NB & LR**
	Precision	**0.99**	**0.98**	**0.89**	**0.63**	
	Recall	**0.74**	**0.74**	**0.71**	**0.61**	
	F1_score	**0.85**	**0.85**	**0.79**	**0.62**	
Food	**Accuracy**	**0.90**	**0.90**	**0.90**	**0.85**	**NB & LR**
	Precision	**1.00**	**1.00**	**1.00**	**0.87**	
	Recall	**0.90**	**0.90**	**0.90**	**0.85**	
	F1_score	**0.95**	**0.95**	**0.95**	**0.86**	
Service	**Accuracy**	**0.84**	**0.84**	**0.83**	**0.73**	**NB & LR**
	Precision	**1.00**	**0.99**	**0.97**	**0.74**	
	Recall	**0.84**	**0.84**	**0.83**	**0.73**	
	F1_score	**0.91**	**0.91**	**0.89**	**0.73**	
Comfort	**Accuracy**	**0.49**	**0.52**	**0.44**	**0.40**	**NB**
	Precision	**0.79**	**0.97**	**0.54**	**0.41**	
	Recall	**0.49**	**0.52**	**0.44**	**0.40**	
	F1_score	**0.59**	**0.68**	**0.48**	**0.40**	
Price	**Accuracy**	**0.62**	**0.64**	**0.55**	**0.48**	**NB**
	Precision	**0.95**	**1.00**	**0.68**	**0.50**	
	Recall	**0.62**	**0.64**	**0.55**	**0.48**	
	F1_score	**0.75**	**0.77**	**0.60**	**0.49**	
Facilities	**Accuracy**	**0.63**	**0.64**	**0.60**	**0.50**	**NB**
	Precision	**0.94**	**0.99**	**0.85**	**0.51**	
	Recall	**0.63**	**0.64**	**0.60**	**0.50**	
	F1_score	**0.50**	**0.77**	**0.70**	**0.49**	

Table 3 Comparison of classifiers

VII. Conclusion And Futurework

In the last decade, hotel industry has witnessed enormous amount of reviews provided by Customers. Plenty of research has been done on analysing this data to extract meaningful and actionable information. This area still is in nascent stage owing to complexities in deciphering human sentiments and intensity of the words used. There are several topic level sentiment analysis tools available in the market, however they are limited to providing document level bias of the feedback and fail to identify and extract implicitly mentioned sentiments. In this approach, we have used LDA algorithm to identify topics out of Hotel reviews dataset and successfully extracted user opinions on particular subset. Also, coherence and perplexity are used to find the optimal number of topics and analyse sentiments to present an easy to understand outcome for users.

Overall, Naïve Bayes gave highest accuracy compared to other classifiers for each topic and the sentiment analysis has provided accuracy up to 90% in few sub-groups. However, this method is semi-automatic because LDA topics are manually labelled based on the weightage of the words generated from LDA.

Further work is suggested to increase the accuracy of all topics (sub-groups) of the model using classifiers and sentiment analysis. Further research is also recommended to completely automate the labelling of data which may be important while dealing with large amount of data. In this approach we have used Machine learning technique but would recommend a deep learning technique to be evaluated to compare both the models.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, 1955. “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, 1892. A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, pp.68–73.

[3] Girolami M and Kaban A 2003 On an equivalence between PLSI and LDA Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03 433

[4] Lu Y, Mei Q and Zhai C 2010 Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA Information Retrieval 14 178-203 R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. Journal of Machine Learning Research,pp 993- 1022.

[6] McCallum, A. 2002. MALLET: A Machine Learning for Language Toolkit. Retrieve from: http://mallet.cs.umass.edu..

[7] M. Hu and B. Liu, 2004. “Mining and Summarizing Customer Reviews,” in Proc. of 10th ACM SIGKDD international Conf. on Knowledge Discovery and Data Mining, pp. 168–177,

[8] B. Lu, M. Ott, C, Cardie, and B. K. Tsou, 2011 “Multi-Aspect Sentiment Analysis with Topic Models,” in Proc. of the 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 81–88,

[9] D. Weinshall, D. Hanukaev, and G. Levi, 2013. “LDA topic model with soft assignment of descriptors to words,” in Proc. of the 30th International Conference on Machine Learning (ICML-13),

[10] Makki, S. Brooks and E. E. Milios, 2014. “Context-Specific Sentiment Lexicon Expansion via Minimal User Interaction,” in Proc. of the International Conference on Information Visualization Theory and Applications (IVAPP),pp. 178–186

[11] Chen, Y., Yu, B., Zhang, X. and Yu, Y., 2020. Topic Modeling For Evaluating Students' Reflective Writin g.

[12] J. Boyd-Graber and P. Resnik, 2010. “Holistic sentiment analysis across languages: multilingual supervised latent Dirichlet allocation,” in Proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 45–55,

[13] F. Li, M. Huang, and X. Zhu, 2010. “Sentiment Analysis with Global Topics and Local Dependency,” In Proc. of AAAI, pp. 1371–1376,

[14] Chen, Y., Yu, B., Zhang, X. and Yu, Y., 2020. Topic Modeling For Evaluating Students' Reflective Writing.

[15] R. Feldman, B. Rosenfeld, R. Lazar, J. Livnat dan and B. Segal, 2006. "Computerized Retrieval and Classiflcation: An application to Reasons", *Intelligent Data Analysis*, vol. 10, no. 2, pp. 183-186,

[16] R. B. Yates dan and B. R. Neto, 1999. Modern Information Retrieval, Harlow:Addison Wesley,

[17] S.S., R. and Dr.P., P., 2018. Topic Categorization on Social Network Using Latent Dirichlet Allocation

[18] ZHANG, Z., MIAO, D. and GAO, C., 2013. Short text classification using latent Dirichlet allocation

[19] Gundla, A., 2016. A Review on Sentiment Analysis and Visualization of Customer Reviews. International Journal Of Engineering And Computer Science,.

[20] D. M. Blei, Apr. 2012. ‘‘Probabilistic topic models,’’ Commun. ACM, vol. 55, no. 4, pp. 77–84,

[21] Saf21.eu. 2020. [online] Available at: <http://www.saf21.eu/wp-content/uploads/2017/09/5004a165.pdf> [Accessed 2 September 2020] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, 2017, pp. 165-174, doi: 10.1109/DSAA.2017.61.

[22] B. Liu, 2009. “Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing,” Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA,

[23] Citeseer.ist.psu.edu. 2020. *Mallet: A Machine Learning For Language Toolkit.*. [online] Available at: <http://citeseer.ist.psu.edu/showciting?cid=573264> [Accessed 11 July 2020]. McCallum, A. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

[24] Jones, Z. and Wallach, H., 2016. Inference on the Effects of Observed Features in Latent Space Models for Networks. *SSRN Electronic Journal*,.

[25] Grubert, E., 2016. Implicit prioritization in life cycle assessment: text mining and detecting metapatterns in the literature. *The International Journal of Life Cycle Assessment*, 22(2), pp.148-158.

[26] Blei D, Lafferty J 2006. Correlated topic models. Adv Neural Inf Proces Syst 18:147

[27] Akhtar, N., Zubair, N., Kumar, A. and Ahmad, T., 2017. Aspect based Sentiment Oriented Summarization of Hotel Reviews. *Procedia Computer Science*, 115, pp.563-571

[28] Douven, I. and Meijs, W., 2007. Measuring coherence. Synthese, 156(3), pp.405-425.

[29] Lau, J., Baldwin, T. and Newman, D., 2013. On collocations and topic models. ACM Transactions on Speech and Language Processing

[30] Suaysom, N. and Gu, W., 2018. Expert Opinion and Coherence Based Topic Modeling. *SSRN Electronic Journal*,.

[31] D. Newman, E.V. Bonilla, and W. Buntine. 2011. Improving topic coherence with regularized topic models. In Advances in Neural Information Processing Systems 24, pages 496–504.

[32] Ray, S., Ahmad, A. and Kumar, C., 2019. Review and Implementation of Topic Modeling in Hindi. *Applied Artificial Intelligence*, 33(11), pp.979-1007.

[33] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, pages 262–272

[34] H X Shi and X J Li 2011."A sentiment analysis model for hotel reviews based on supervised learning," in in International Conference on Machine Learning and Cybernetics China

[35] C Jingnian H Huang S Tian and Y Qu 2009. "Feature selection for text classification with Naïve Bayes," Expert Systems with Applications: An International Journal pp 5432-5435

[36] H X Shi and X J Li 2011 "A sentiment analysis model for hotel reviews based on supervised learning," in in International Conference on Machine Learning and Cybernetics China

[37] "6 Tren Wisata Utama Tahun 2016," tripadvisor 14 December 2015 [Online] Available: https://www.tripadvisor.co.id/TripAdvisorInsights/w665 [Accessed 3 September 2020]

[38] C Jingnian H Huang S Tian and Y Qu 2009."Feature selection for text classification with Naïve Bayes," Expert Systems with Applications: An International Journal pp 5432-5435

[39] Brownlee, J., 2020. *A Gentle Introduction To The Bag-Of-Words Model*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/> [Accessed 17 August 2020].

[40] Farisi, A., Sibaroni, Y. and Faraby, S., 2019. Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *Journal of Physics: Conference Series*, 1192, p.012024.

[41] R.K, Mishra., Urolagin, S. and Joshi, J., 2019.'A Sentiment Analysis-Based Hotel Recommendation Using TF-IDF Approach'. pp.pp.811–815..

[42] D. Chen, C. S. Ong, and L. Xie, 2016 ."Learning Points and Routes to Recommend Trajectories," pp. 2227–2232,

[43] O. Claveria and S. Torra, 2017 ."The appraisal of machine learning techniques for tourism demand forecasting," no. March, pp. 59–89,

[44] M, M., 2020. [online] Mran.microsoft.com. Available at:<https://mran.microsoft.com/web/packages/naivebayes/naivebayes.pdf> [Accessed 19 August 2020] version 0.9.6 https://CRAN.R-project.org/package=naivebayes

[45] Rokach, L. and Maimon, O., n.d. Data Mining With Decision Trees. 3rd ed. Singapore: World Scientific Publishing 2015. 305 pp

[46] Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine Learning, 20(3), pp.273-297.

[47] Gholami, R. and Fakhari, N., 2009. Support Vector Machine: Principles, Parameters, And Applications.

[48] P. Samui, S. Sekhar, V.E. Balas (Eds.), (2017), Handbook of Neural Computation, Elsevier, San Diego pp. 515-535

[49] MARILL, K., 2004. Advanced statistics: linear regression,*1Part I: simple linear regression. *Academic Emergency Medicine*, 11(1), pp.87-93.

[50] Guo, Y., Barnes, S. J., & Jia, Q. 2017. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. TOURISM MANAGEMENT, 59, 467-483.

[51] Alaei, A., Becken, S. and Stantic, B., 2017. Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, 58(2), pp.175-191

[52] Serra Cantallops, A. and Salvi, F., 2014. New consumer behavior: A review of research on eWOM and hotels. International Journal of Hospitality Management, 36, pp.41-51

[53] Appel O, Chicalana F, Carter J, Fujita H 2016 .A hybrid approach to the sentiment analysis problem at the sentence level. Knowl Based Syst 108:110–124

[54] Pranckevičius, T. and Marcinkevičius, V., 2017. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, 5(2).

[55] ResearchGate. 2012. *(PDF) Classification Of Customer Reviews Based On Sentiment Analysis*. [online] Available at: <https://www.researchgate.net/publication/252067764_Classification_of_Customer_Reviews_based_on_Sentiment_Analysis> [Accessed 12 August 2020].

[56] Blei, David, Andrew Ng, and Michael Jordan. 2003 "Latent Dirichlet Allocation." Journal of Machine Learning Research. pp. 993-1022

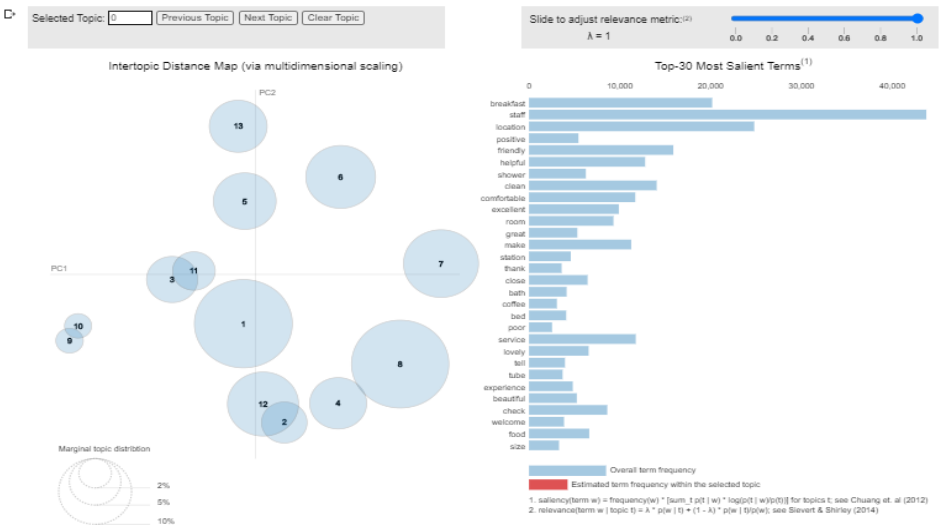
APPENDICES

i. Output generated by LDA

	Document_No	Topic_Num	Topic_Perc_Contrib	Keywords	Review_Title	dominent_topic_Label
0	181618	0.0	0.9750	shower, night, window, floor, work, bath, open, door, could, location	Room warm even though ground floor turned down thermostats couldn open windows both partner kept...	Facilities
1	9898	1.0	0.9667	staff, amazing, would, view, really, love, recommend, beautiful, location, feel	second visit this Hotel expected were treated wonderful Hotel standards very high departments no...	Staff
2	349727	2.0	0.9654	positive, star, staff, need, clean, location, dirty, poor, smell, work	Internet extremely poor constantly slow weak This only reason makes never want this place again ...	Staff, Facilities
3	6602	3.0	0.9500	desk, staff, help, front, always, reception, restaurant, dinner, people, smile	visited here work charity dinner amazing staff dealt with were professional 500+ meals achieve...	Service,Staff
4	137252	4.0	0.9763	back, free, come, leave, also, would, staff, lovely, make, upgrade	booked nights this weekend wedding Although wedding ceremony reception held staff still upgraded...	Location,Staff
5	234820	5.0	0.9690	breakfast, coffee, location, clean, comfortable, room, choice, staff, could, excellent	King actually singles pushed together didn feel level Wifi weak Cooked breakfast items bread cof...	Food
6	878	6.0	0.9500	staff, location, friendly, helpful, clean, excellent, breakfast, great, comfortable, good	Stayed here nights. Great location. Easy walking distance London Eye, Covent Garden, Trafalgar S...	Cleanliness, COmfort , Facilities
7	234502	7.0	0.9743	station, close, location, walk, clean, minute, restaurant, breakfast, train, parking	heavy view from wasn paid basic rates this expected Would definitely again Good size bath with f...	Location
8	511462	8.0	0.9678	service, check, food, staff, restaurant, wait, drink, slow, long, order	Overall performance staff seemed pretty poor Ordered lunch waited very long when went chase offe...	Service , Food
9	296775	9.0	0.9678	book, night, staff, check, give, tell, ask, would, double, pay	Booked three which wasnt available when arrived Fortunately worker twin able accommodate third p...	Staff,Location

ii. Word-Cloud of topics

iii. Topic visualisation of LDA through pyLDAvis Gensim



iv. TSNE Representation of topics.