Customer Reviews Analysis Platform by Correlating Sentiment Analysis and Text Clustering

Ehtisham ur Rehman[1], Najam Aziz[2], Nasir Ahmad[2]

[1]Dept. of Computer Science, University of Engineering and Technology (UET), Peshawar, Pakistan

[2]Dept. of Computer Systems Engineering, University of Engineering and Technology (UET),

***Correspondence**: ehtishamrehman@uetpeshawar.edu.pk, najamkhattak@gmail.com, n.ahmad@uetpeshawar.edu.pk

Citation | Rehman. E, Aziz. N, Ahmad. N, "Customer Reviews Analysis Platform by Correlating Sentiment Analysis and Text Clustering", IJIST, Special Issue. pp 312-328, June 2024

Received | May 15, 2024, **Revised |** May 22, 2024 **Accepted |** May 30, 2024 **Published |** June 03, 2024.

Customer reviews and feedback are of paramount importance in the improvement cycle of any industry, product, or service. Formerly, product ratings were the basis for performance evaluation and key drivers of improvements. However, ratings were unable to depict the complete picture and were not adequate for an in-depth analysis of any product or service. Hence, customer reviews become the ultimate source of providing feedback for a specific detailed analysis as well as contributing to performance metrics. Although, customer reviews provide a very essential measure for performance evaluation, extracting important features and topics from customer reviews has been challenging due to its unlabeled and variant nature. This paper focuses on extracting topics from customer review data and bringing in use the of implicit knowledge for analytics. To extract topics and clusters from review data, unsupervised machine learning algorithms such as K-Means and Latent Dirichlet Allocation (LDA) are used. These topics are then correlated with sentiment analysis - score of positive or negative feedback - of each customer review. The products or services are then categorized with the help of the topics or domains they belong to alongside the sentiments. This provides a valuable analysis such as the score of positive, neutral, and negative feedback for each customer review input to new customers as well as product managers. This research work aims to use the hotel reviews dataset to categorize and rank hotels based on the different services captured in the text from customer reviews. The research work makes use of the hotel reviews dataset for categorizing and ranking hotels based on the different services discussed in the customer's reviews text. Moreover, this paper also provides a visualization of both text clustering algorithms depicting the topics in each cluster for an insightful analysis.

Keywords: Text Clustering, K-means, Latent Dirichlet Allocation Algorithm, Sentiment Analysis, Customer Reviews Feedback.

Introduction:

The rise of e-commerce and online shopping platforms has exponentially enhanced the importance of customer reviews and feedback on these platforms. These reviews are not only a source of guidance for new potential buyers but also provide valuable input to the manufacturers or service providers for analyzing customer demands for new product designs as well as improving existing products and services [1] Customer feedback is gathered in various forms on an online platform, it can be either in the form of ratings on a numeric scale or answering a certain set of questions about a specific product or service [2]. However, the most common is the written reviews from the customers and are most widely used for analysis among business analysts for assessing the performance due to the different dimensions of the product discussed by customers in their reviews.

With this upsurge in customer reviews and feedback, different statistical and sentimental analysis approaches have been adopted to unfold the hidden insights in the data [3]. Sentimental Analysis focuses on extracting the underlying opinion within the text that can range from positive to negative values thus helping in identifying the sentiment and customer satisfaction about the product or service [4]. On the other hand, statistically, these reviews can help in identifying the most sold products in different quarters of the year as well as people's interest in different products geographically. Thus, paving the way for different recommendation systems based on these statistics alongside prediction and machine learning algorithms [5].

Besides sentiment analysis, customer reviews also provide an opportunity to detect the dominant topics or dimensions discussed in these review texts. These topics can be related to the price, quality, shipping, or any other service related to the product. However, due to the unlabeled nature of the review datasets, unsupervised learning becomes the only option to identify dominant topics. Identifying dominant topics in the reviews will enable the manufacturers or service providers to specifically focus on those areas of concern for performance improvement. For this purpose, different unsupervised topic modeling techniques are available to cluster similar data in common groups and identify the most discussed topics in these groups [6].

In this research work, the aforementioned topic modeling techniques are combined with sentiment analysis to not only determine the opinion of the customer but also correlate it with the dominant category the reviewer is referring to. Moreover, the research work also compares the two topic modeling approaches commonly used, LDA and K-Mean, for better categorization of reviews. In addition, the paper evaluates the performance of the aforementioned two clustering algorithms through the evaluation matrix. Furthermore, the paper also provides a visualization of the identified dominant topics in both the clustering algorithms to enable a better comparison of the two techniques. In addition, sentiment analysis and topic modeling become the stepping-stone for conducting statistical analysis in the results of Section 4 such as ranking products and services based on different topics alongside the review sentiment score.

The remaining Section of the paper is structured in the following way. Section 2 presents the related literature review as well as the background knowledge required to conduct the research. Section 3 of the paper describes the use case of the European hotel's dataset as well as the methodology applied to conduct the research. Section 4 discusses the results and graphs obtained by using different statistical analysis approaches as well as evaluating the two topic modeling techniques based on their performance matrices. Finally, Section 5 concludes the research work with the help of the results obtained as well as illustrating future research work prospects in this domain.

Literature Review:

Related Work:

Due to the utter importance of customer feedback, there has been an increasing focus among industries on analyzing and classifying customer reviews for attaining a competitive advantage over competitors in respective domains. Several research works have focused on extracting the opinion by applying sentiment analysis techniques while others have focused on using topic modeling techniques. Jeong et al. [7] categorized reviews from Moscow's hotels into five specific attributes. They focus on several attributes such as amenities, experience, location, transactions, and value by applying regression analysis, etc. However, the research work did not extract dominant topics from the customer reviews to provide a more in-depth analysis to the hotel owners.

In [8], Siew Hoong et al. have crawled review data of four-star and five-star hotels from Kuala Lumpur. This work focused on finding the most used predominant themes such as location, restaurant, comfort, etc. The topics discussed in the reviews are extracted through SAS text miner using the R language. However, the paper did not relate it to the sentiment described in each review rather it just finds the topic in each review. Authors have applied the Context-Based Keyword Pattern Cluster Analysis (CBKPC) using R-Mini to cluster similar topics together.

Moreover, in [9], Tian et al. analyzed the review dataset from four-star and five-star hotels from four cities in China. This work focused on finding the sentiment score i.e. positive, negative, and neutral as well as finding the customers' interest in categories related to food, staff, services, etc. They use natural language processing to clean text by removing stop words and unwanted words. Then it manually extracted categories by reviewing small parts of the dataset and applied sentiment analysis to the dataset. The authors also presented a correlation between the categories and frequency between the categories. However, the research work only focused on applying sentiment analysis to customer reviews. The research work in [9] is limited by the fact that it manually extracts categories that are not possible for the complete dataset.

Furthermore, in [10] Porntrakoon and Moemeng coupled sentiment analysis with different dimensions in the customer reviews text. The research work focused on analyzing the sentiment or polarity of the three pre-defined dimensions in the reviews which are pricing, product, and shipping by using multi-dimensional lexicon and sentiment compensation techniques. Although the framework designed in [10] finds the polarity of three different dimensions of the text of the reviews, it discards any other dimension except the predefined ones and is thus unable to analyze any unseen topic or dimension in the review text.

Text Clustering and Topic Modelling Algorithms:

Topic modeling has been widely used in numerous domains ranging from text clustering to collaborative filtering, from information retrieval to image analysis [11], and relation extraction to infer hidden insights from the data [12]. The topic Model Algorithm usually mines out the most dominant topics and semantic words from the text document [13]. Among the extracted topics similar ones are then clustered into the same groups. On the other hand, text clustering has a wide range of applications such as organizing and browsing documents, finding a coherent summary of the text document collection, and document classification [14].

Although, topic modeling is often studied as a separate research area from text clustering, and it majorly focuses on determining latent topics from the text it is one of the most widely used methods for probabilistic document clustering [15]. Probabilistic Latent Semantic Indexing (PLSI) and LDA are the two frequently used methods for topic modeling via probabilistic document clustering. The PLSI and LDA differ by the way term-document and topic-document probabilities are modeled. The probabilistic document clustering aims at creating a probabilistic generative model for the entire text document. Each document in probabilistic document clustering belongs to one of the k topics. The topic modeling algorithm usually mines out the most dominant topics and semantic words from the text document [16].

Besides the topic modeling algorithm, Fasheng Liu and Lu Xiong have analyzed and divided text clustering algorithms into different categories: hierarchical clustering, distance-based partitioned clustering, density-based algorithm, self-organizing maps algorithm on the basis of several factors such as scalability, dimensionality, dependence on input parameters and ability to deal noise [17]. In this research work, LDA, a topic modeling method from the probabilistic document clustering algorithms, and K-means, a distance-based text clustering algorithm are implemented and evaluated based on their respective evaluation scores.

Clustering:

Clustering techniques are one of the most widely recognized data analyses in the machine learning area and are used to get an instinct about the structure of the data. They can be defined as the task of identifying subgroups in the data such that data points in the same subgroups (cluster) are fundamentally the same while data points in different clusters are different [18]. Particularly, we attempt to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similar measure, for example, Euclidean-based or relationship-based separation. The decision of which similarity measure to utilize is application-specific.

K- Mean Algorithm:

The K-mean algorithm initially arbitrarily assigns a specified k number of cluster centers in space [18]. Afterward, each sample data point is assigned to these centers based on the nearest Euclidean distance between the data point and the cluster centers [19]. Then iteratively the center is recomputed from the mean of all the samples in the respective cluster and the process of assigning data points to the cluster is repeated until the cluster centers no longer move significantly [20]. The K-mean algorithm implementation cycle is depicted in Figure 1. The implementation cycle starts with data collection and the much-needed text pre-processing.

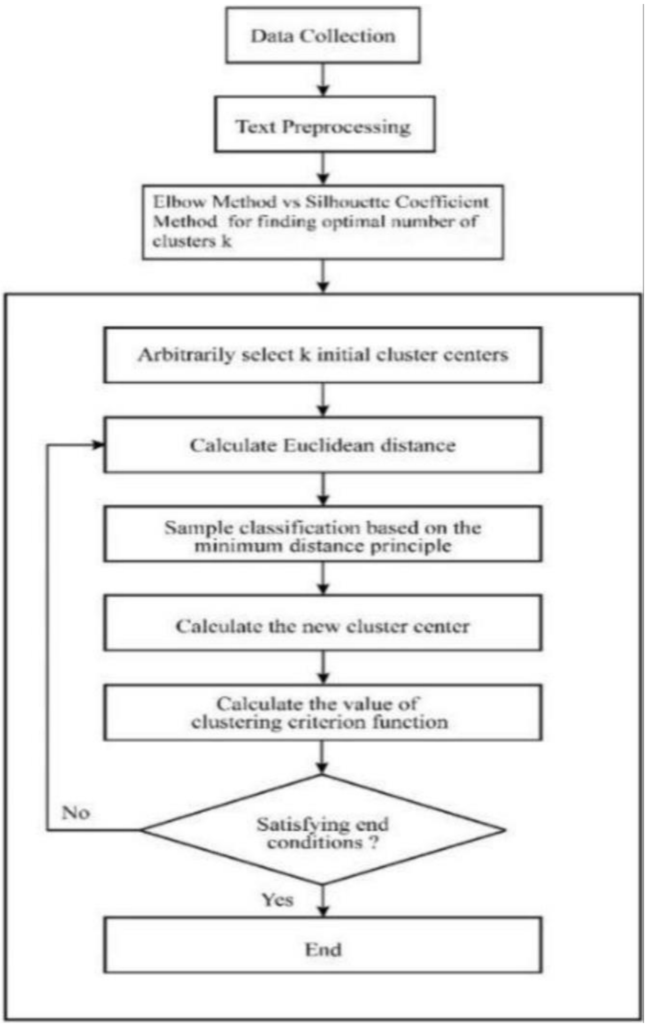

Data Collection
Text Preprocessing
Elbow Method vs Silhouette Coefficient Method for finding optimal number of clusters k
Arbitrarily select k initial cluster centers
Calculate Euclidean distance
Sample classification based on the minimum distance principle
Calculate the new cluster center
Calculate the value of clustering criterion function
Satisfying end conditions ?
No
Yes
End

Figure 1: K-mean algorithm implementation cycle

The way the K-means algorithm works is as follows:

- First, determine the number of clusters k.
- Initialize centroids by first rearranging the datasets and then randomly selecting k data points for the centroids without substitution.
- Continue the repeating process until there is no change to the centroids i.e. the assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.
- Refer each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of all data points that belong to each cluster.

The goal of the k-means algorithm is to reduce the objective function, which is called the squared error function. Equation 1 gives the objective function (J).

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} ||x_i^{(j)} - C_j||^2 \quad (1)$$

The K-mean algorithm distributes the n number of cases into k number of clusters as shown in Figure 2 that are predefined. Whereas Euclidean distance between a case and the centroid. However, to get the appropriate number of K for the optimized implementation of the K-mean algorithm silhouette method is used [21].

The K-mean algorithm distributes the n number of cases into k number of clusters as shown in Figure 2 that are predefined. Whereas $||^{()} - ||^2$ is an Euclidean distance between $^{()}$ and the centroid. However, to get the appropriate number of K for the optimized implementation of the K-mean algorithm silhouette method is used [21].

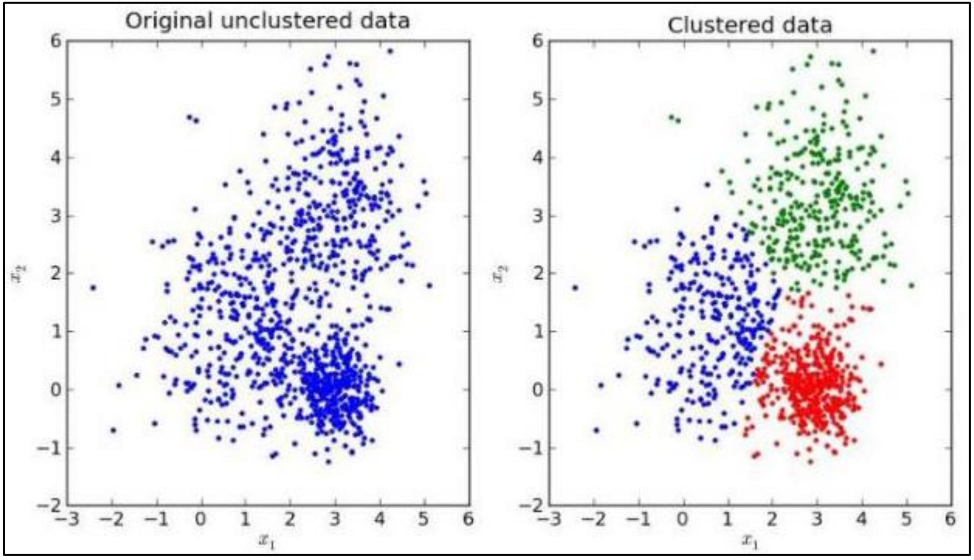


Figure 2: K-mean data distribution in K number of clusters

Latent Dirichlet Allocation:

In natural language, processing, latent Dirichlet allocation is a generative factual model that enables sets of perceptions to be clarified by unobserved groups that clarify why a few parts of the data/information are comparable. Topic modeling using LDA (Latent Dirichlet allocation) allows us to extract the hidden topic themes from a large dataset, using LDA a mixture of topics from a large number of documents can be found [22][23]. A defined number of topics is assumed within the document or set of reviews by LDA [24]. LDA enables to process of the frequency of each topic based on the occurrence of it in the dataset [22]. LDA extracts the hidden structure of topics from the data by applying a probabilistic approach. Latent Dirichlet Allocation works on the Bayesian estimation framework to find out the theme topics from the dataset [7]. Latent Dirichlet Allocation is the unsupervised probabilistic model that takes a bag of words i.e. corpus dictionary as input [25]. LDA filters out topics from the document and then finds words for topics. The working of LDA is given in Equation 2.

$$p(w|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n=1}^{k} p(z_n|\theta)p(w_n|z_{n,}\beta)\right)d\theta$$

The main idea of LDA is, that the documents are expressed as random mixtures over latent topics, where each topic is categorized by distribution over words [26]. Equation 2 comprises several parameters. The k is the number of topics, and are corpus-level parameters and the variables are document-level variables. The variables are word-level variables and are sampled once for each word in each document. LDA involves three levels of evaluation, through which the topic node is sampled repeatedly within the document. Under the LDA model, multiple topics associated with the document are established.

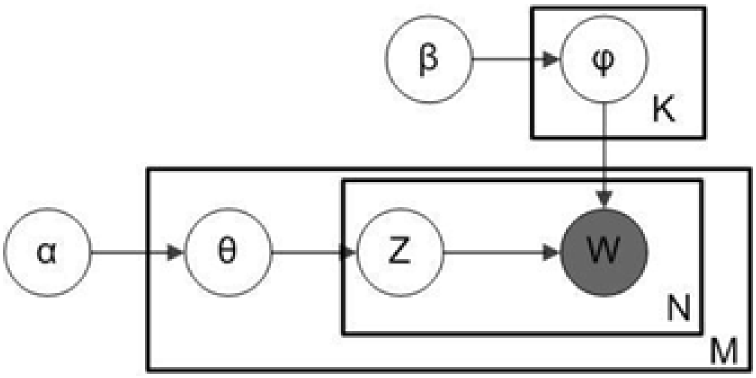


Figure 3: LDA (Latent Dirichlet Allocation)

LDA groups the words in a document based on the frequency and its association with the topics [25][27]. Once the LDA model is trained on data, dominant topics are extracted and categorized. The use case of hotel review data is a model using LDA and the dominant topics are categorized such as hotel services, staff, transactions, location, and restaurant.

Sentiment Analysis:

The sentiment is an emotion or attitude prompted by the feelings of the customer. Sentiment analysis is also known as opinion mining, which evaluates people's opinions towards any topic, product, or service. Sentiment analysis is an emerging domain in the area of research in natural language processing (NLP) and has gained attention in recent years in text mining and text classification. It is a machine-learning approach to analyze and classify the user's comments, emotions, opinions, and attitudes based on the polarity of the text. Sentiment analysis plays an important role in depicting and finding people's opinions on politics, e-businesses, and the general social trend toward an ongoing issue. In politics, sentiment analysis is used for forecasting the outcomes of political trends in the region and predicting poll results [25]. Whereas in business it is widely used to analyze and predict stock market trends. Sentiment Analysis has gained more importance in online applications, e-commerce, and social media platforms i.e. blogs, Twitter, online website stores, and discussion forums which attracts customers, organization firms, and stakeholders to do analysis on data from online websites (2) and to extract meaningful information from the data [28][29]. The main aim of sentiment analysis is to analyze user opinions and classify whether they fall into which category i.e. negative sentiment, positive sentiment, or Neutral sentiment.

Implementation:

This section describes the implementation process of text clustering algorithms, LDA, and K-means, alongside sentiment analysis to conduct in-depth analysis. The implementation will be executed in the steps shown in Figure 4. The approach used is generic and is applicable for the analysis of any review's dataset in any business domain. This paper analyses the European Hotels review dataset.

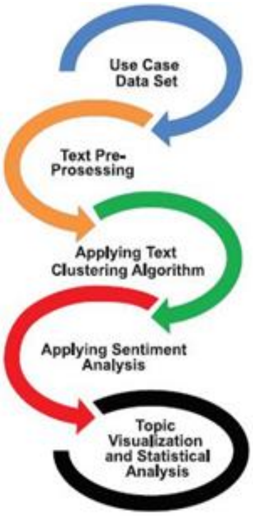
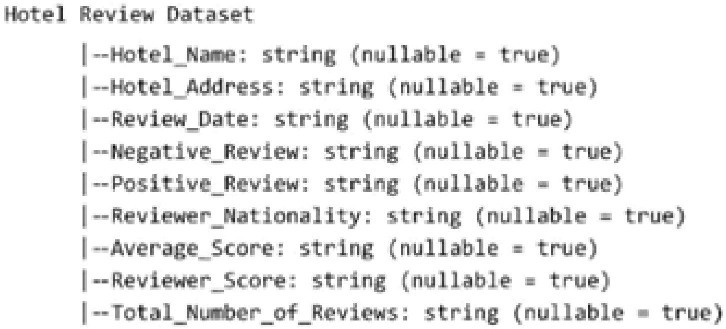


Figure 4: Implementation Execution Steps

Use Case Dataset:

The dataset contains 515k review records of 13 European hotels and has been sourced from an online website of datasets called Kaggle. The dataset contains several attributes as shown in Figure 5 such as hotel address, hotel name, review date, average review score, the total number of reviews, reviewer nationality, negative review, positive review, etc. Each of the customer reviews provides a textual description of their personal opinion or experience and ratings regarding the hotel services acquired by the customer. The review text in the dataset is unlabeled, thus unsupervised text clustering algorithms are applied to categorize the reviews. The review text is mostly unclean and has a lot of unwanted words that are removed in the data preprocessing stage.



```
Hotel Review Dataset

      |--Hotel_Name: string (nullable = true)
      |--Hotel_Address: string (nullable = true)
      |--Review_Date: string (nullable = true)
      |--Negative_Review: string (nullable = true)
      |--Positive_Review: string (nullable = true)
      |--Reviewer_Nationality: string (nullable = true)
      |--Average_Score: string (nullable = true)
      |--Reviewer_Score: string (nullable = true)
      |--Total_Number_of_Reviews: string (nullable = true)
```

Figure 5: Available Attributes Column in the Dataset

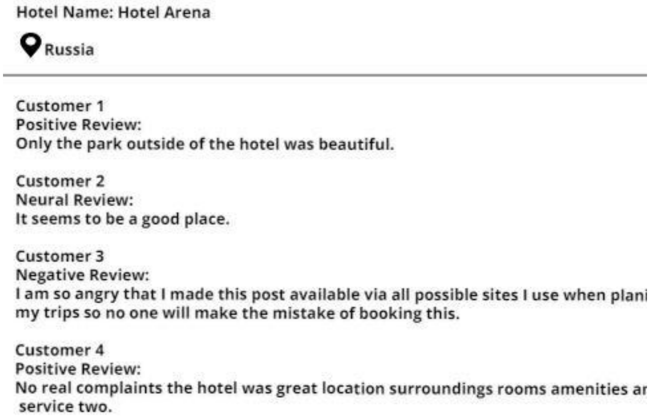
Text Preprocessing:

Text Pre-Processing is the conventional method used for natural language processing tasks. It performs the transformation of text into a more digestible form so that the machine learning algorithms can use it for better performance and magnificent efficiency. The raw data is pre-processed before applying the machine-learning algorithm. The pre-processing includes cleaning data using natural language techniques, such as tokenization, and removing words that have fewer than 3 characters. Moreover, after removing unwanted words, the remaining words are lemmatized and stemmed.

Text Pre-Processing:

Tokenization is a process of chopping long sentences or text streams into words and phrases by removing unwanted characters such as empty spaces and punctuations. Every token is made of a word from the first character to the last character [30].

In this research, we take customer reviews i.e. positive, negative, and neural as raw data for tokenization as shown in Figure 6**,** in which we convert the customer reviews into individual words because the Python programming language doesn't understand any distinction between words, they are just a stream of characters which are all same.



Hotel Name: Hotel Arena

Russia

Customer 1
Positive Review:
Only the park outside of the hotel was beautiful.

Customer 2
Neural Review:
It seems to be a good place.

Customer 3
Negative Review:
I am so angry that I made this post available via all possible sites I use when plani
my trips so no one will make the mistake of booking this.

Customer 4
Positive Review:
No real complaints the hotel was great location surroundings rooms amenities ar
service two.

Figure 6: Different Customers Review Data

Removal Stop Words:

The review text data consist of excessively common words such as "of", "be", "but", "on", and other common words. Although these stop words have a very small weight due to their frequent occurrences these words have a very high frequency [30].

We only take the important keywords from the desired topics i.e. comfort, staff, location rooms, etc. as shown in Table 1, and convey the context that is present in the review. Therefore, these "I am", and "In there" are present more frequently in many reviews and these are just connectors, and there doesn't add too much to the context that the machine wants to understand. In this procedure, we get rid of these stop words or punctuation used by the customers in their reviews.

Stemming and Lemmatization:

Although stemming and lemmatization are used for transforming words to their original base form both these methods work differently [31]. Stemming tries to achieve it by removing the suffixes and prefixes, but that may sometimes lead to incorrect word forms. So to counter the limitation posed by stemming, lemmatization is done which takes into account the morphological analysis of the words [31]. However, lemmatization needs a dictionary, which it can consult to connect the form of the word to its root. After the text, pre-processing the dictionary is created from pre-processed data. A Corpus dictionary is created from the pre-processed dictionary, which reports the frequency of words appearing in the dictionary showing the number of occurrences of a word in each document.

Determine the Optimal Number of K For K-Mean, Silhouette Score, and LDA:

K-Means:

K-means is an unsupervised machine-learning algorithm for clustering large datasets. K-means groups the data into different clusters based on Euclidean distance. The optimal number of the cluster for the K-Mean algorithm is evaluated using the Elbow method. Although a manual selection of several clusters k is accurate it comes with extensive computation overhead to analyze each k. On the contrary, the elbow method is employed to find the optimal number of clusters for the K-means algorithm by varying k stepwise i.e. starting k at 1 to 10, and calculating the accuracy and computational cost on the training of the model [32]. The Elbow method computes the total sum of square error within the cluster in each iteration [33][34]. After computing the sum of squares, the elbow method plots the curve according to the number of clusters k. The blend knee plot will show the optimal number of clusters k as shown in Figure 7.

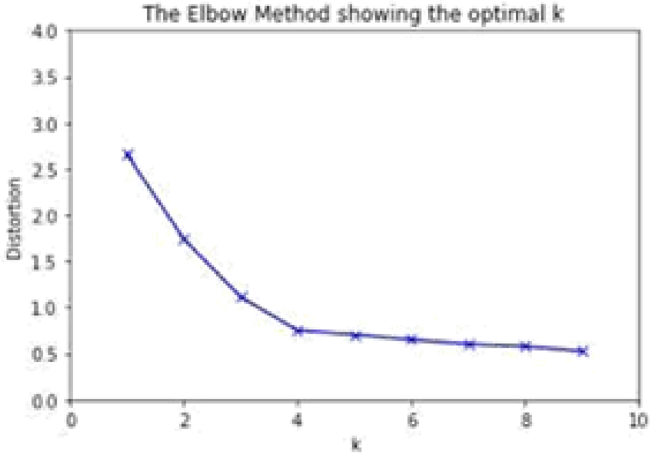


Figure 7: Elbow Method to find optimal K for the K-mean algorithm

Silhouette Score:

Silhouette is the better measure to choose the number of clusters to be defined from the data [35]. It is determined for each instance and the Equation 3 goes like this:

$$\boldsymbol{Silhouetee\ Coefficient} = (\boldsymbol{a} - \boldsymbol{b}) / \mathbf{max}(\boldsymbol{a}, \boldsymbol{b}) \qquad \mathbf{(3)}$$

Where b is the mean intra-cluster distance: mean distance to the other instances in the same cluster. a depicts the mean closest cluster distance i.e. mean distance to the instance of the following nearest cluster. The coefficient varies between -1 and 1. A value near 1 implies that the instance is near to its cluster and is a part of the correct cluster. While a value near -1 means that the value is allocated to the inappropriate cluster.

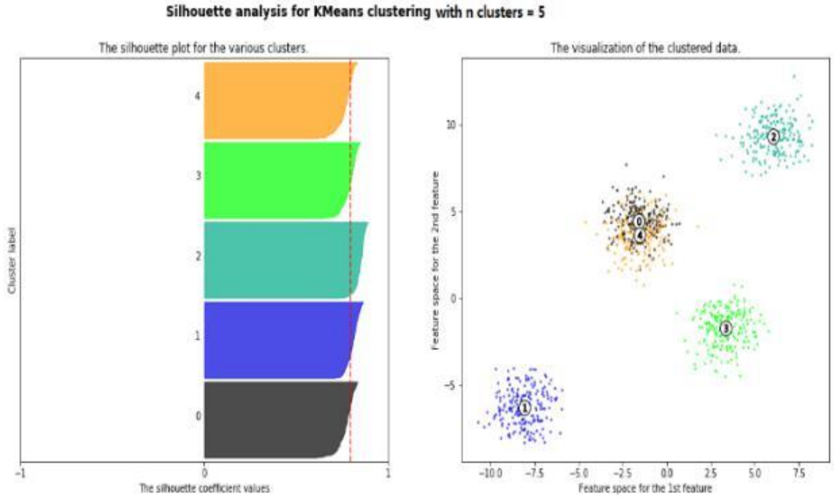


Figure 8: Silhoutter Method

According to the strategy, k=5 should be chosen for the number of clusters. This method is better as it decides the optimal number of clusters more significant and clearer. However this metric is computationally expensive as the coefficient is calculated for every instance. Therefore, the decision regarding the optimal metric to be chosen for the number of cluster decisions is to be made according to the requirements of the product.

In Figure 8, the range of n clusters is 2 to 11, and the thickness of the silhouette plot of the cluster size can be visualized. The silhouette plot for cluster 0 when n clusters are equal to 2, is bigger owing to the grouping of the 3 sub-clusters into one big cluster. However, when the n clusters are equal to 5, all the plots have a similar thickness. The average silhouette score with several clusters is given below in Figure 8, in which at cluster 5 the graph bends at 0.7958870 value.

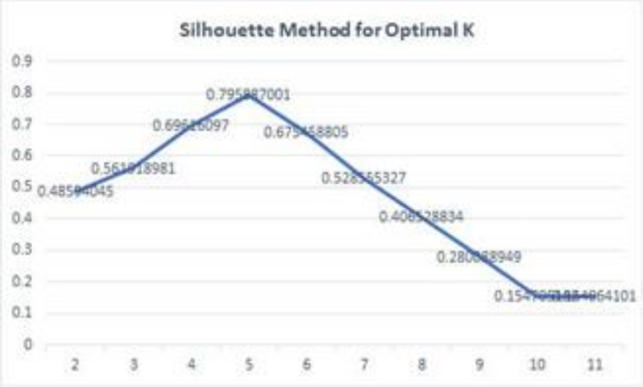


Figure 9: Average Silhouette Score

For implementing the LDA algorithm, the value of optimal k was determined using the log-likelihood and perplexity scores. The higher the log-likelihood score and the smaller the perplexity score, the better the model [36]. To find the highest value of log-likelihood and smallest value of perplexity, a different value of k with a different set of learning decays was tested. The learning decay rate defines the change in weights during training. The learning rate controls how quickly the model is adapted, and thus for large learning decays the changes are more rapid whereas for smaller learning rates the changes are smaller with each update and thus require more training epochs in comparison to large learning rates. The optimal value of k is determined from six different values of k, which are 5, 10, 15, 20, 25, and 30, by varying three different values of learning decay i.e. 0.5, 0.7, and 0.9. Figure 9 depicts the optimal value for k=5 and learning decay

= 0.7 when the log-likelihood has the largest value i.e. - 19578767.67.

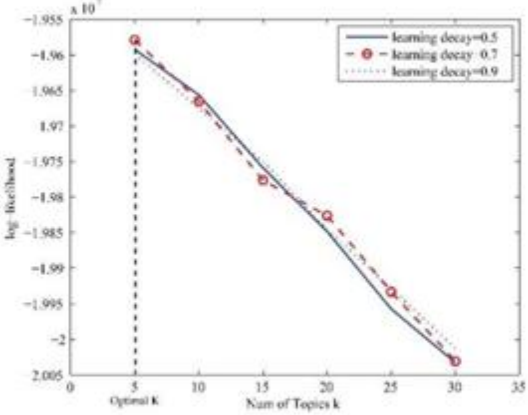


Figure 10: Log-likelihood vs Num of topics graph for finding optimal value of k for LDA algorithm

Sentiment Analysis via TextBlob:

TextBlob is a Python library used for textual processing of raw data and can be used efficiently for sentiment analysis. TextBlob allows us to perform common operations on textual data along with natural language processing tasks such as Noun phase extraction, parsing, n-grams, tokenization, classification, sentiment analysis, and much more. NLP operations are accessed by calling the TextBlob API [37]. In TextBlob sentiment analysis property called sentiment, returns two properties polarity and subjectivity of text. Polarity has its score range which starts from [-1.0,1.0] float score, with -1 being extremely negative and 1 being extremely positive. Reviews, whose sentiment score is less than zero are termed negative reviews, while reviews that have sentiment scores greater than 0.25 are labeled as positive reviews. Moreover, reviews that fall in the range of 0 to 0.25 are referred to as neutral reviews. Subjectivity has also a range of float scores from [0,1] refers to personal opinion emotion or factual information.

Results:

This Section illustrates and visualizes the list of topics and keywords extracted from the reviews with the help of machine learning algorithms. The Section also discusses results obtained from performing statistical analysis on the use case dataset of European hotels with the help of topic categorization and sentiment analysis. Moreover, in this Section, the two topic modeling algorithms are also evaluated based on their respective performance metrics.

Extracted Topics and Keywords:

The topic-modeling algorithm mentioned in the implementation Section created five different clusters with a different set of keywords. The topics discussed in these five clusters differ from each other in the context inferred from the keywords. The top 10 keywords with the highest frequency in each of the clusters were selected to infer the context and to label each of the topics. The five different topics inferred from the keywords are comfort and transaction, staff behavior and services, location and accessibility, room and hotel facilities and decor, and finally, food and restaurants. Table 1 shows the top 10 keywords in each of the topic clusters along with their labels.

Table 1: Most important keywords extracted for each topic

Comfort and Transaction (Topic 1)	Staff Behavior and Service (Topic 2)	Location and Accessibility (Topic 3)	Room, Hotel Facilities, and Décor (Topic 4)	Food and Restaurants (Topic 5)
comfort	staff	location	room	breakfast
hotel	friendly	position	clean	food
upgrade	service	station	View	restaurant
room	clean	close	specious	coffee
free	helpful	metro	bathroom	buffet
extra	breakfast	city center	modern	service
timing	polite	tram	décor	quality
check-in/out	professional	transport	pool	delicious
Welcome	Welcome	park	terrace	price
stay	comfort	access	bed	staff

The distance between each topic and the frequency of each word in the topic can best be visualized with LDAvis [38], a visualization library available in Python. LDAvis is a web-based interface specifically designed for LDA algorithm topic visualization. It aims to provide a better understanding and deep inspection of the topics generated from the LDA algorithm by illustrating the meaning of each topic, the prevalence of each topic, and the correlation between different topics. In Error! Reference source not found., a visualization of a European hotel dataset with five different clusters is shown, which is produced through the LDAvis library. The visualization in Figure 10 is divided into two parts, the left panel of the visualization shows the five clusters with five circles plotted in a two-dimensional plane. The distance between the centers represents the inter-topic distance thus showing the correlation between different topics. The area of each cluster circle represents the prevalence of that cluster topic, the larger the area of a cluster circle the more is its prevalence, and vice versa [36].

The second part of the visualization in Figure 10 is the right panel with a horizontal bar chart of individual terms. The bar charts represent the most relevant terms and frequency of each term in the selected topic on the left panel thus helping in interpreting the meaning of each topic. The bar charts show the top 30 terms of each topic extracted with the help of LDA and then ranked based on estimated frequency and relevance in the selected topic thus providing a clear picture of each topic and its correlation with other topics.

Besides the LDA topic modeling algorithm, the K-means algorithm is also used to cluster the data. Figure 11 shows the different clusters made by the K-means algorithm. The optimal number of k is determined with the help of the Silhouette method as described in

Implementation Section 3. Although attributes like simplicity and fast convergence make the K-means algorithm a viable option for clustering data, the initial arbitrarily assigned cluster centers significantly hamper its performance [39]. Due to this random selection, the algorithm may converge to locally optimal solutions. Therefore, this paper makes use of the topics extracted via the LDA algorithm for statistical analysis in the next Section due to its more realistic results than k-mean for the topic assignment.

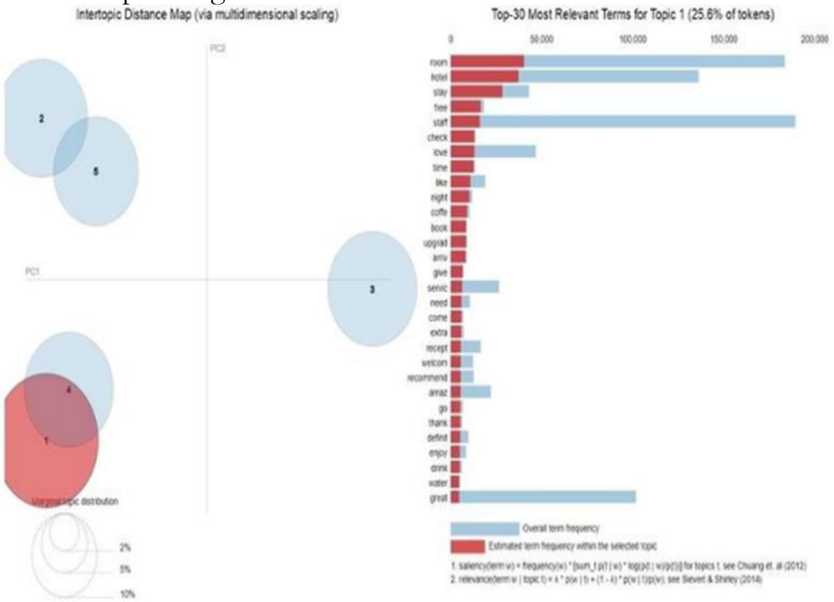


Figure 11: LDA topics Cluster

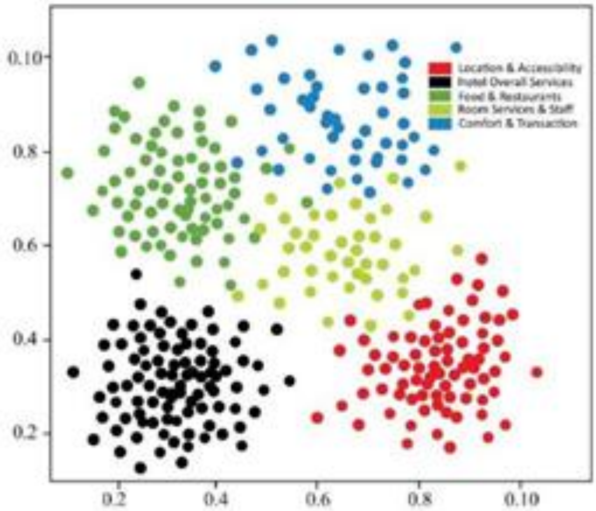


Figure 12: K-means cluster with five different sets of cluster data

Statistical Analysis of Use Case Data Set:

After applying the text clustering algorithm on the use case dataset of the European Hotels reviews, five different topics were extracted from the data that became the basis for statistical analysis alongside the sentiment analysis score calculated for each review in this Section. In total, the data set contains 515738 reviews of hotel customers from different parts of Europe. These 515738 reviews are categorized into five groups as shown in Figure 14, which are then further divided based on sentiment score into three subgroups namely: positive, negative, and neutral reviews. The sentiment score is calculated with the help of the text blob library available in Python as mentioned in the implementation Section 3 for each review. The total set of reviews is divided into the following five groups and subgroups with their respective statistics.

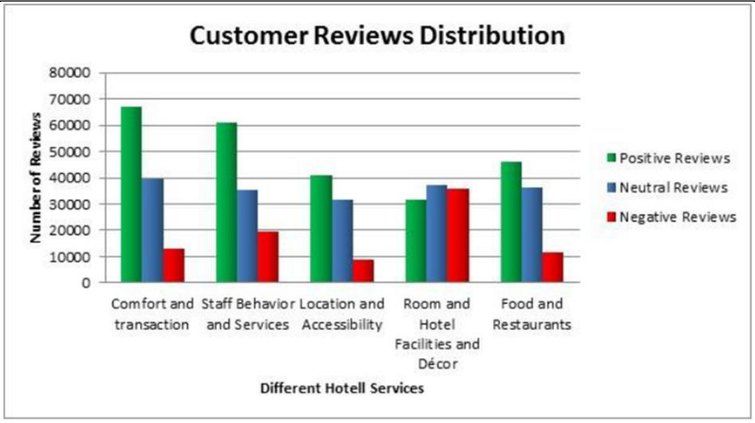


Figure 13: Customer Reviews distribution based on different hotel service

Comfort and Transaction:

Comfort and Transaction topics discuss comfortability provided by the hotel in issues such as bookings, payments, check-in/check-out times, complimentary breakfast, and room upgradation. There are 119658 (23.2%) reviews that discuss the overall services of the rather than one specific topic or service of the hotel. Among these 67017 are positive reviews, 39433 have sentiment scores in the neutral range and 13208 are negative reviews as illustrated in Figure 12.

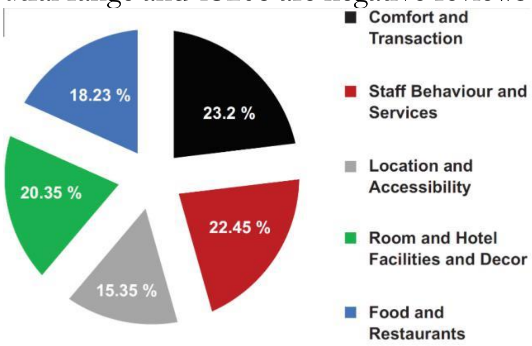


Figure 14: Number of Reviews in Percentages for each topic category

Staff Behavior and Service:

The Staff Behavior and Service cluster consists of 115807 (22.45%) reviews. This topic encompasses all the staff behaviors and services-related issues that include staff professionalism and friendliness, staff behavior, staff readiness, and response time. As Figure 12 shows, 31646 reviews fall in the positive category, 35452 in the neutral region and 11706 have sentiment scores less than 0 thus belonging to the negative reviews group.

Location and Accessibility:

This group contains around 81323, which account for 15.77% of the total reviews. The group comprises the reviews in which the users have discussed the location and surrounding facilities available such as environment, accessibility from bus stops and airports, location of the nearby markets, historical sites, city center, and restaurants. Among the 81323 reviews, 40774 have sentiment scores greater than 0.25, hence they are termed as positive reviews. Whereas, 31585 reviews are neutral and 8964 people have negative opinions about the surrounding facilities of the hotels.

Room and Hotel Facilities and Décor:

The Room and hotel Facilities and decor cluster highlight the facilities and interior of the room as well as the overall hotel. The terms discussed in this group include room decor, view from the room, floor, and roof design, availability of pool and terrace area in the hotel, and the spaciousness of rooms and bathrooms. There are 104945 (20.35%) reviews that discuss the overall services of the rather than one specific topic or service of the hotel. Among these 31646 are positive reviews, 37358 have a sentiment score in the neutral range and 35941 are negative reviews as depicted in Figure 12.

Food and Restaurants:

This cluster of reviews has 94005 (18.23%) reviews, which majorly discuss the food services available such as breakfast, buffet, and drinks. Moreover, it talks about the quality and different variety of foods and drinks (coffee and tea) served in the restaurants inside the hotels as well as the distance from the famous food outlets nearby the hotel they reside in. Across the total reviews belonging to this cluster, 45914 reviews are positive opinions about the food and restaurants, while 36385 reviews have sentiment scores between 0 and 0.25 thus termed as neutral and 16989 reviews belong to the negative category.

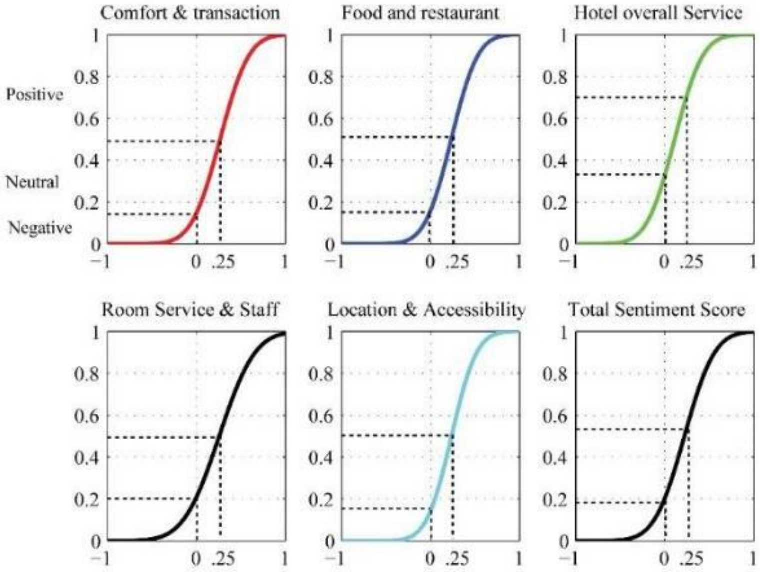


Figure 15: CDF plots of each service's sentiment score distribution

Figure 14 is the illustration of the distribution of the sentiment score of reviews for each topic through Cumulative Distribution Function (CDF) graphs. The CDF plots presented in Figure 14 represent the empirical cumulative distribution function of the sentiment score. The sentiment labels (positive, neutral, and negative) on the y-axis depict the percentage of customer sentiments for each topic. The ranges are defined by the dotted line in Figure 14. CDF plots are useful for comparing the distribution of different ranges of data.

The graphs are marked with the threshold of each category namely; Positive, neutral, and negative sentiment scores. The first five graphs show the distribution of sentiment scores of each topic generated from the topic modeling algorithms. The last graph is the distribution of combined scores of all the customer reviews. The paper aims to provide a visualization tool for customers, product managers, and business decision-makers to extract topics from any customer review data and visualize the corresponding sentiments and topics using the CDF plots.

Conclusion:

This paper provided a novel approach to sentiment analysis based on topic extraction. This helps in utilizing implicit knowledge for analytics and useful decision-making. Unsupervised machine learning algorithms such as K-Means and Latent Dirichlet Allocation (LDA) for clustering and topic modeling were employed. The proposed approach has the potential to be a valuable analytic tool for new customers as well as product managers. The research work used the hotel reviews dataset as a use case to evaluate the proposed approach. The hotel review dataset was categorized and ranked hotels based on the different services discussed in the customer reviews text.

In the future, we aim to provide a more general framework by leveraging advanced machine learning tools such as deep learning to the review's dataset. This will help in extracting topics from the review's dataset including hotels, airlines, places to visit, clinics, hospitals, and many more that will potentially assist customers and travelers in useful decision-making.

Nomenclatures	
( )	Case
	Centroid
	Document-level variables
	Word level variables
J	Objective function
K	Optimized implementation of K-mean algorithm
k	Number of clusters
n	Range of clusters
Greek Symbols	
	Corpus level parameters
	Corpus level parameters
Abbreviations	
JESTE	Journal of Engineering Science and Technology
C	
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
CDF	Cumulative Distribution Function
PLSI	Probabilistic Latent Semantic Indexing
CBKP	Context-Based Keyword Pattern Cluster Analysis
C	National Center in Big Data and Cloud Computing
NCBC	

References:

[1] Z. Singla, S. Randhawa, and S. Jain, "Statistical and sentiment analysis of consumer product reviews," 8th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2017, Dec. 2017, doi: 10.1109/ICCCNT.2017.8203960.

[2] K. L. Santhosh Kumar, J. Desai, and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," 2016 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2016, May 2017, doi: 10.1109/ICCIC.2016.7919584.

[3] H. Zhang, A. Sekhari, F. Fourli-Kartsouni, Y. Ouzrout, and A. Bouras, "Customer reviews analysis based on information extraction approaches," IFIP Adv. Inf. Commun. Technol., vol. 467, pp. 227–237, 2016, doi: 10.1007/978-3-319-33111-9_21/TABLES/4.

[4] Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," Proc. 2016 IEEE Int. Conf. Big Data Anal. ICBDA 2016, Jul. 2016, doi: 10.1109/ICBDA.2016.7509790.

[5] Y. Saito and V. Klyuev, "Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes," Int. Conf. Adv. Commun. Technol. ICACT, vol. 2019-February, pp. 681–685, Apr. 2019, doi: 10.23919/ICACT.2019.8702039.

[6] S. Shivashankar, S. P. Algur, and P. S. Hiremath, "Cluster Analysis of Customer Reviews Extracted from Web Pages," J. Appl. Comput. Sci. Math., vol. 4, no. 9, pp. 56–62, Jan. 2010, Accessed: Jun. 08, 2024. [Online]. Available: https://doaj.org/article/73234c83e1d3441d96e355982b5a0eb0

[7] "Text Analytics of Online Customer Reviews." Accessed: Jun. 08, 2024. [Online]. Available: https://ecommons.cornell.edu/server/api/core/bitstreams/49fa121a-80cf-4126-bffc-ecde99faede0/content

[8] A. S. . Lee, Z. Yusoff, Z. Zainol, and P. V, "Know your Hotels Well! an Online Review Analysis using Text Analytics," Int. J. Eng. Technol., vol. 7, no. 4.31, pp. 341–347, Dec. 2018, doi: 10.14419/IJET.V7I4.31.23406.

[9] X. Tian, W. He, R. Tao, and V. Akula, "Mining Online Hotel Reviews: A Case Study from Hotels in China".

[10] P. Porntrakoon and C. Moemeng, "Thai sentiment analysis for consumer's review in multiple dimensions using sentiment compensation technique (SenSecomp)," ECTI-CON 2018 - 15th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol., pp. 25–28, Jul. 2018, doi: 10.1109/ECTICON.2018.8619892.

[11] T. Iwata, T. Hirao, and N. Ueda, "Topic Models for Unsupervised Cluster Matching," IEEE Trans. Knowl. Data Eng., vol. 30, no. 4, pp. 786–795, Apr. 2018, doi: 10.1109/TKDE.2017.2778720.

[12] M. Allahyari and K. Kochut, "Discovering Coherent Topics with Entity Topic Models," Proc. - 2016 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2016, pp. 26–33, Jan. 2017, doi: 10.1109/WI.2016.0015.

[13] B. Wang, Y. Liu, Z. Liu, M. Li, and M. Qi, "Topic selection in latent dirichlet allocation," 2014 11th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2014, pp. 756–760, Dec. 2014, doi: 10.1109/FSKD.2014.6980931.

[14] R. Bekkerman, R. El-Yaniv, Y. Winter, and N. Tishby, "On feature distributional clustering for text categorization," SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval), pp. 146–153, 2001, doi: 10.1145/383952.383976.

[15] "Mining Text Data." Accessed: Jun. 08, 2024. [Online]. Available: https://sci-hub.se/10.1007/978-1-4614-3223-4

[16] M. Alhawarat and M. Hegazi, "Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents," IEEE Access, vol. 6, pp. 42740–42749, Jul. 2018, doi: 10.1109/ACCESS.2018.2852648.

[17] F. Liu and L. Xiong, "Survey on text clustering algorithm: Research present situation of text clustering algorithm," ICSESS 2011 - Proc. 2011 IEEE 2nd Int. Conf. Softw. Eng. Serv. Sci., pp. 901–904, 2011, doi: 10.1109/ICSESS.2011.5982485.

[18] A. Sudha Ramkumar and R. Nethravathy, "TEXT DOCUMENT CLUSTERING USING K-MEANS ALGORITHM," Int. Res. J. Eng. Technol., p. 1764, 2008, Accessed: Jun. 08, 2024. [Online]. Available: www.irjet.net

[19] C. Xiong, Z. Hua, K. Lv, and X. Li, "An improved K-means text clustering algorithm by optimizing initial cluster centers," Proc. - 2016 7th Int. Conf. Cloud Comput. Big Data, CCBD 2016, pp. 265–268, Jul. 2017, doi: 10.1109/CCBD.2016.059.

[20] S. Lu et al., "Clustering method of raw meal composition based on PCA and kmeans," Chinese Control Conf. CCC, vol. 2018-July, pp. 9007–9010, Oct. 2018, doi: 10.23919/CHICC.2018.8482823.

[21] M. Bertin and I. Atanassova, "K-means and Hierarchical Clustering Method to Improve our Understanding of Citation Contexts".

[22] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," Tour. Manag., vol. 59, pp. 467–483, Apr. 2017, doi: 10.1016/J.TOURMAN.2016.09.009.

[23] A. Kelaiaia and H. F. Merouani, "Clustering with Probabilistic Topic Models on Arabic Texts," Stud. Comput. Intell., vol. 488, pp. 65–74, 2013, doi: 10.1007/978-3-319-00560-7_11.

[24] J. Büschken and G. M. Allenby, "Sentence-Based Text Analysis for Customer Reviews," https://doi.org/10.1287/mksc.2016.0993, vol. 35, no. 6, pp. 953–975, Jul. 2016, doi: 10.1287/MKSC.2016.0993.

[25] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," J.

Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[26] C. Sievert and K. E. Shirley, "LDAvis: A method for visualizing and interpreting topics," pp. 63–70, Jun. 2014, doi: 10.3115/V1/W14-3110.

[27] D. Kozlowski, V. Semeshenko, and A. Molinari, "Latent Dirichlet allocation model for world trade analysis," PLoS One, vol. 16, no. 2, p. e0245393, Feb. 2021, doi: 10.1371/JOURNAL.PONE.0245393.

[28] "Sentiment Analysis on Political Tweets." Accessed: Jun. 08, 2024. [Online]. Available: https://www.researchgate.net/publication/311986158_Sentiment_Analysis_on_Political_Tweets

[29] W. Wang, "Sentiment analysis of online product reviews with semi-supervised topic sentiment mixture model," Proc. - 2010 7th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2010, vol. 5, pp. 2385–2389, 2010, doi: 10.1109/FSKD.2010.5569528.

[30] B. Saberi and S. Saad, "Sentiment analysis or opinion mining: A review," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 5, pp. 1660–1666, 2017, doi: 10.18517/IJASEIT.7.5.2137.

[31] L. M. Abualigah, A. T. Khader, and M. A. Al-Betar, "Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering," Proc. - CSIT 2016 2016 7th Int. Conf. Comput. Sci. Inf. Technol., Aug. 2016, doi: 10.1109/CSIT.2016.7549453.

[32] P. Han, S. Shen, D. Wang, and Y. Liu, "The influence of word normalization in English document clustering," CSAE 2012 - Proceedings, 2012 IEEE Int. Conf. Comput. Sci. Autom. Eng., vol. 2, pp. 116–120, 2012, doi: 10.1109/CSAE.2012.6272740.

[33] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," Int. J. Comput. Appl., vol. 105, no. 9, pp. 17–24, 2014, doi: 10.5120/18405-9674.

[34] S. Tripathi, A. Bhardwaj, and P. E, "Approaches to Clustering in Customer Segmentation," Int. J. Eng. Technol., vol. 7, no. 3.12, pp. 802–807, Jul. 2018, doi: 10.14419/IJET.V7I3.12.16505.

[35] V. Divya and K. N. Devi, "An Efficient Approach to Determine Number of Clusters Using Principal Component Analysis," Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018, Nov. 2018, doi: 10.1109/ICCTCT.2018.8551182.

[36] H. Jelodar et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimed. Tools Appl., vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/S11042-018-6894-4/METRICS.

[37] Z. J. Wang, D. Choi, S. Xu, and D. Yang, "Putting Humans in the Natural Language Processing Loop: A Survey," Bridg. Human-Computer Interact. Nat. Lang. Process. HCINLP 2021 - Proc. 1st Work., pp. 47–52, Mar. 2021, Accessed: Jun. 08, 2024. [Online]. Available: https://arxiv.org/abs/2103.04044v1

[38] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018, pp. 533–538, Nov. 2018, doi: 10.1109/ISEMANTIC.2018.8549751.

[39] "Silhouette Texture", [Online]. Available: https://svbrdf.github.io/publications/siltex/siltex.pdf